THEORY AND DECISION LIBRARY / VOLUME 22

TEDDY SEIDENFELD

PHILOSOPHICAL PROBLEMS OF STATISTICAL INFERENCE

Learning from R. A. Fisher

D. REIDEL PUBLISHING COMPANY
DORDRECHT : HOLLAND / BOSTON : U.S.A. / LONDON : ENGLAND

PHILOSOPHICAL PROBLEMS OF STATISTICAL INFERENCE

THEORY AND DECISION LIBRARY

AN INTERNATIONAL SERIES
IN THE PHILOSOPHY AND METHODOLOGY OF THE
SOCIAL AND BEHAVIORAL SCIENCES

Editors

GERALD EBERLEIN, *University of Technology, Munich*
WERNER LEINFELLNER, *University of Nebraska*

Editorial Advisory Board:

K. BORCH, *Norwegian School of Economics and Business Administration*
M. BUNGE, *McGill University*
J. S. COLEMAN, *University of Chicago*
W. KROEBER-RIEL, *University of Saarland*
A. RAPOPORT, *University of Toronto*
F. SCHICK, *Rutgers University*
A. SEN, *University of London*
W. STEGMÜLLER, *University of Munich*
K. SZANIAWSKI, *University of Warsaw*
L. TONDL, *Prague*
A. TVERSKY, *Stanford University*

VOLUME 22

TEDDY SEIDENFELD
University of Pittsburgh

PHILOSOPHICAL PROBLEMS OF STATISTICAL INFERENCE

Learning from R.A. Fisher

D. REIDEL PUBLISHING COMPANY
DORDRECHT:HOLLAND/BOSTON:U.S.A.
LONDON:ENGLAND

Library of Congress Cataloging in Publication Data

Seidenfeld, Teddy.
Philosophical problems of statistical inference.

(Theory and decision library; v. 22)
Bibliography: p.
Includes index.
1. Mathematical statistics–Philosophy.
2. Fisher, Ronald Aylmer, Sir, 1890–1962. I. Title.
QA276.S43 519.5 79–11488
ISBN 90–277–0965–3

Published by D. Reidel Publishing Company,
P.O. Box 17, Dordrecht, Holland

Sold and distributed in the U.S.A., Canada, and Mexico
by D. Reidel Publishing Company, Inc.
Lincoln Building, 160 Old Derby Street, Hingham,
Mass. 02043, U.S.A.

All Rights Reserved
Copyright © 1979 by D. Reidel Publishing Company, Dordrecht, Holland
No part of the material protected by this copyright notice may be reproduced or
utilized in any form or by any means, electronic or mechanical,
including photocopying, recording or by any informational storage and
retrieval system, without written permission from the copyright owner

Printed in The Netherlands

To Janice

TABLE OF CONTENTS

PREFACE xi

CHAPTER 1/PROBABILITY AND INVERSE INFERENCE

1.1. Introduction 1
1.2. Bayes' Theorem and the Laplacean Principle of Insufficient Reason 5
1.3. Inductive Relevance and Related Concepts: Likelihood, Sufficiency and Ancillarity 16
1.4. Continuity: The Cumulative Distribution Function and Density Function 21

CHAPTER 2/NEYMAN-PEARSON THEORY

2.1. Introduction to the Neyman-Pearson Theory of Hypothesis Testing and Confidence Intervals 29
2.2. Hypothesis Testing 37
2.3. Two Uses of Randomizers to Produce 'Mixed' Tests with Improved Power over 'Pure' Tests 42
2.4. Tests with Composite Alternative Hypotheses 45
2.5. Hacking's Criticism of N-P Hypothesis Testing 47
2.6. The Sampling Rule Problem: Another Consequence of the Forward Look 49
2.7. Confidence Intervals 51
2.8. Composite Confidence Intervals and Confidence Equivalence Condition 58
2.9. An Example Illustrating the Use of Randomization for Constructing Confidence Intervals 63

CHAPTER 3/FISHERIAN SIGNIFICANCE TESTING

3.1. Introduction to the Fisherian Theories of Significance Testing and Fiducial Probability 71
3.2. A Reconstruction of Fisherian Significance Tests 76

3.3. Data too Good to be True 84
3.4. Illegitimate Transformations of the Random Variable 86
3.5. Mistaken Analyses of Relevance and the r.v. 90
3.6. Composite Null Hypotheses, Ancillarity and Repeated Sampling 95

CHAPTER 4/THE FIDUCIAL ARGUMENT: ONE PARAMETER

4.1. Introduction to the Fiducial Argument 105
4.2. The Pivotal Variable and the Smooth Invertibility Condition 107
4.3. Ancillarity, Smooth Invertibility and the CC Principle 112
4.4. Case 1: The Fiducial Argument with Data of Two Kinds 116
4.5. Case 2: The Fiducial Argument with Data of One Kind and a Common Sufficient Statistic 120
4.6. Case 2: Lindley's Paradox and Canonical Pivotal Variables 124
4.7. Two Notes on the Connection Between Significance and Fiducial Probability 128
4.8. Case 3: Data of One Kind without a Common Sufficient Statistic, a Paradox and the Principle of Insufficient Reason 129

CHAPTER 5/THE FIDUCIAL ARGUMENT: SEVERAL PARAMETERS

5.1. Multivariate Fiducial Probability and the Step-by-Step Procedure 137
5.2. Multivariate Fiducial Probability and Repeated Sampling 146
5.3. Additional Factorizations and the Five Parameter Bivariate Normal 149
5.4. Review of Paradoxes: (1) Conflicts with N-P Theory 151
5.5. Review of Paradoxes: (2) Conflicts with Bayesian Theory 155

CHAPTER 6/IAN HACKING'S THEORY

6.1. Ian Hacking's Likelihood Principle of Irrelevance and the Fiducial Argument 168
6.2. Hacking's Principle of Irrelevance and Generalized Inverse Inference 178

CHAPTER 7/HENRY KYBURG'S THEORY

7.1. Henry Kyburg's 'Rationally Representative' Predicate and the Fiducial Argument 183

7.2. Kyburg's Relevance Conditions and Randomization 195

CHAPTER 8/RELEVANCE AND EXPERIMENTAL DESIGN

8.1. Fisherian Theories of Relevance and Experimental Design 202
8.2. Randomization and Relevance 214
8.3. Summary 219

APPENDICES

9.1. Pivotal Probabilities 226
9.2. The Equivalence of Fisher's and Tukey's Pivotal Conditions 227
9.3. Coherence of One Parameter Fiducial Arguments 228
9.4. Extended Coherence of Fiducial Probability 229
9.5. Hacking's Irrelevance Principle and Tukey's Smooth Invertibility Requirement 230
9.6. Hacking's Second Frequency Principle 231
9.7. Notes for the Bivariate Normal Distribution, Including a Derivation of the Master Significance Test [MST] 232
9.8. Additional Notes for Fieller's (also, the MST) Interval Estimate of the Ratio of Means in a Bivariate Normal Distribution 236

BIBLIOGRAPHY 239

INDEX 243

PREFACE

Examine the table of contents of a recent issue of a major statistics journal and you will be impressed by the diversity of categorized topics, applied and theoretical. This small sample accurately reflects the discipline's wide horizon. But you are mistaken if you conclude that the detailed categorization is evidence of some grand statistical theory that covers the expanse of even one journal number. Statistics is fragmented into a host of different approaches that, on occasion, vie for reception of their analyses.

It is my position in this book that, from a philosophical point of view, the most important competition takes place within the arena of *inverse inference*: statistical inference from an observed 'sample' to an unobserved 'population'. Problems of inverse inference are distinguished because they reveal the fundamental differences across alternative approaches and because they are basic problems in applied statistics.

The principal object of this work is a reconstruction and evaluation of Fisherian statistics, with special attention to Fisher's idea of fiducial probability as it pertains to inverse inference. In the first Chapter I review several postulates of inductive logic that serve in later chapters as the conditions of adequacy for my account of the Fisherian program. Both as motivation for the discussion of Fisherian methods and as illustration of the content of the postulates I offer, in Chapter 2, criticisms of 'orthodox statistics', the approach due to J. Neyman and E. S. Pearson.

The next three chapters provide my version of Fisher's theory of significance testing and fiducial probability. I model my account on the connection (found in the 'orthodox' position) between hypothesis testing and confidence intervals. In Chapter 3 I discuss significance testing. In Chapter 4 I claim to find and expose the fatal flaw in fiducial inference, a flaw that is distinct from the many published objections to the fiducial argument. Seriatim, I deal with these problems in Chapter 5 and show how they may be solved. (The published objections involve, for the most part, multivariate fiducial probability. Thus, the first sections of Chapter 5 reconstruct fiducial inference with several parameters.)

Chapters 6 and 7 examine, respectively, the original theories of Ian Hacking and Henry Kyburg. Each offers his own account of fiducial

probability. I contrast their versions of Fisher's argument with my own and show where their reconstructions falter.

A second theme that I develop is the claim for the existence of a tie between methodological considerations pertinent to inverse inference (such as randomization in the design of experiments) and inferential considerations (such as principles of relevance in the fiducial argument). The second parts of Chapters 6 and 7 broach the issue, thereby fixing another point of contrast between Hacking's and Kyburg's positions. The main force of this analysis appears in the final chapter where I suggest a theory of experimental design that unifies questions of design with questions of inductive relevance. The concluding section identifies the conflict that exists in one's attitudes towards inductive relevance when the justification of fiducial inference is compared to the justification of randomization.

(Those readers seeking additional comments on the organization of my arguments in this book can find them in Section 1 of Chapter 1. Also, each chapter is headed by a synopsis of its contents. These synopses may be read to give a more detailed overview of the book, and to guide those readers who desire a selection of specific issues.)

It is with great pleasure that I publicly acknowledge some of my outstanding intellectual debts. I am grateful beyond measure to my teacher, Isaac Levi, for having introduced me to the principal subject of this book, fiducial probability, and for having offered philosophic guidance throughout my graduate studies at Columbia University. My understanding of the basic issues in inductive logic reflects his clear and incisive thought, which he freely shares.

I have had the honor of knowing Henry Kyburg since my undergraduate days. His numerous, detailed comments have been a stimulus for me to rethink difficult questions. Most important of all, the manner in which he approaches his own work serves as my model for philosophic criticism.

I thank Sidney Morgenbesser for his many suggestions and for his patience when I have been slow to act on them. Jonathan Lieberson has contributed his friendship and keen advice with equal generosity. Also, I have greatly benefited from discussion with each of the following scholars: Alberto Coffa, J. K. Ghosh, Ian Hacking, Bruce Levin, Charles Parsons, Carl Posy, Bimal Sinha, Burton Singer and Samuel Wieand.

This work grows out of my doctoral dissertation, 'The Fiducial Argument', completed at Columbia University in 1975. Preliminary research, as well as revision of the manuscript during the summer of 1976, was carried out

under the sponsorship of the Woodrow Wilson Foundation. Apart from several references added to recognize important collateral results and not including updated introductory passages, no substantive changes have been made since that revision.

Department of Philosophy
University of Pittsburgh

TEDDY SEIDENFELD
February, 1979

CHAPTER 1

PROBABILITY AND INVERSE INFERENCE

The book opens with a brief introduction indicating the major problems dealt with in this work, namely *direct* and *inverse* statistical inference. My principal goal is established: to reconstruct and evaluate Fisherian inference (1.1). This is followed by a review of an inductive logic, generated from five postulates which serve as conditions of adequacy for all proposed reconstructions. The first three, *coherence, conditionalization* and *total evidence* parallel equivalent Carnapian inductive principles and form the heart of *Bayesian* theory. The fourth, the *Frequency* Principle, incorporates stochastic (probability$_2$) concepts into the inductive logic. The fifth requirement, *objectivity*, assures the uniqueness of confirmational relations between hypotheses and their evidential basis. *Objectivity* is found to be the source of controversial proposals for solutions to inverse inference (1.2).

In anticipation of the technical aspects of the upcoming investigation, two sections (1.3 and 1.4) provide a review of centrally important notions. In 1.3 relevance considerations, especially *sufficiency* and *ancillarity*, are presented. The text explores their connections to the five adequacy conditions. In 1.4 the reader finds a summary of the mathematical machinery to be used for analysis of continuously variable states. *Cumulative distribution functions* (and their *density function* representations) are discussed. Also, in connection with suggestions for inverse inference, Section 1.4 considers *improper* densities and their relation to a Laplacean principle of *Insufficient Reason*. Finally, two (related) density forms, *location* and *scale* parameters, are defined. These play an important role in my reconstruction of Fisher's *fiducial* inference, discussed in Chapters 4 and 5.

1.1 INTRODUCTION

The years 1915 through 1940 represent an eventful era for contemporary mathematical statistics. During this period the foundations were set for what are currently the received positions. These alternative programs, although displaying much variety along several dimensions, can be divided into two broad categories. One family, going by the name 'Bayesians,' traces its roots back to the 18th-century mathematician Thomas Bayes, and includes amongst its members such notables as Laplace, De Morgan, Karl Pearson, J.M. Keynes, Sir Harold Jeffreys, Ruldolf Carnap, B. deFinetti, L. Savage, I.J. Good, and D.V. Lindley. The rival group, loosly known as 'Frequentists,' has developed and flowered mostly in reaction to the Bayesian tradition. Its distinguished membership includes Boole, Venn, R.A. Fisher, E.S. Pearson, J. Neyman, R. von Mises, H. Reichenbach, A. Wald, and noted philosophers I. Hacking and H. Kyburg.

The nature of the disagreement between the two schools is best captured by a comparison of views on two fundamental problems; to wit, 'direct inference' and 'inverse inference.' With the exception of certain radical Bayesians, e.g. deFinetti,[1] it is generally conceded that there exist situations or events incorporating a 'chance' or 'stochastic' ingredient: a property which, however explicated, is independent of the investigator's beliefs and desires. Typical illustrations are: coin flipping, dice rolling, various card games, selecting chips from an urn, counting (by different methods) radioactive emissions, etc. Without going over the details of the argument here, it is claimed that these 'chance' events admit formal representation in the mathematical theory of probability. That is, the 'stochastic' property gives a model for the probability calculus. It is then assumed, both by the Bayesians and Frequentists, that knowing the value of this stochastic probability factor determines the credal 'probability' for an arbitrary or 'random' occurrence of the 'chance' event. This is the *direct* inference.

For example, when flipping a coin, or drawing counters from an urn containing only red and blue chips, the 'stochastic' ingredient is captured in the probability distribution known as the 'binominal distribution'; with parametric factor θ whose values are restricted to the [0, 1] interval. θ represents the 'bias' of the coin (for landing 'heads'), or the proportion of chips which are red. Knowledge of the value of θ, say .5 (which reflects the 'fairness' of the coin, or that the urn contains equal numbers of red and blue counters) determines, by the direct inference principle, the credal probability that a 'flip' of the coin results in 'heads-up,' or that the counter selected from the urn is red. By referring to the general context of the chance event, e.g. the flipping of the coin, or the drawing of a chip from the urn, the direct inference principle can be restated as determining the probability for an arbitrary *trial* of the stochastic set-up resulting in a given outcome. Thus, the probability is .5 that the flip lands heads-up, or that the chip drawn is a red one.

So much is uncontroversial for the majority of Bayesians and Frequentists. However, the two programs diverge sharply for the converse problem, 'inverse inference,' Suppose that the 'bias' of the coin is *unknown*, but that a past record of n flips is available. Is it possible to duplicate the force of the direct inference, so that a conclusion about the value of θ can be expressed 'probabilistically'? The Bayesians answer, "Yes!" The Frequentists retort, "No! At least not unless more information that the track record of n past flips is also available."

As will be discussed below, the Frequentist objections to the Bayesian solution of the inverse inference problem fall into two categories. On the

one hand, it is claimed that their solution is no solution at all, since it recommends blatantly inconsistent results (inconsistent with the probability calculus). Second, where the first difficulty is overcome (through various restrictions designed to insure consistency), the results are claimed to be ad hoc and without the warrant of empirically based regulations.

In lieu of Bayesian inverse inference, the Frequentists adopt a host of alternative statistical procedures, including *significance testing* (Fisher), *hypothesis testing* (Neyman-Pearson), *likelihood analysis* (Fisher), *confidence intervals* (Neyman-Pearson), and *Fiducial probability* (Fisher). Only the last two of these allege to duplicate the Bayesian solution by allowing (in a reduction to direct inference) a 'legitimate' probability involving the potential values of the unknown parameter, given merely the information about the outcomes of previous trials. The principal object of this work is to present a renewed appraisal of R.A. Fisher's program for resolving the inverse inference problem; specifically, an analysis of his most controversial proposal, the fiducial argument.

In order to understand some of the difficulties this program faces, and therefore to motivate a review of Fisher's solutions, consider first the Bayesian position. In the section following, the Bayesian program will be sketched at least so far as to formulate precisely the issues touched upon in this introduction, and to supply clear and meaningful conditions of adequacy for the reconstructed Fisher theory. Even this preliminary grounds clearing must count as 'rational reconstruction,' since Fisher chose to present his views primarily through paradigmatic examples; without either adequately formed citeria for evaluating the results, or precisely stated rules by which the fiducial argument could be derived. Although the first fiducial probabilities were presented as early as 1930,[2] their status with respect to Bayesian alternatives or later Frequentist proposals, e.g. Neyman-Pearson confidence intervals (1934),[3] was not rendered intelligible to the statistical public within Fisher's lifetime.

Despite what must count as a vain effort at both clarification and vindication of the fiducial argument, found in Fisher's last major work *Statistical Methods and Scientific Inference,*[4] the allure of fiducial probability did not compensate for the intractable difficulties it encountered – in terms of scope of application and threats of inconsistency. A sorry comment on the state of fiducial probability, at the late date 1963 (ironically, the year after Fisher's death), is the sentiment expressed by a group of experts gathered at the International Statistical Institute's Conference (on Fiducial Probability): to quote D.V. Lindley, "If this meeting were more formal than it is I would propose the motion that this meeting marks the end of the fiducial

argument."[5] It is the hope of this project to show that Lindley's remarks were premature. A general outline for the endeavor follows.

In the remaining sections of the first chapter the ground is cleared for a reconstruction of inverse inference. The discussion of the Bayesian program clarifies some basic conditions of adequacy, involving the probability calculus and Bayes' Theorem, which are to be imposed on all potential frequentist solutions to the inverse inference problem.

Before developing the Fisherian theory, a chapter (Chapter 2) is devoted to a review of the Nayman-Pearson program which currently stands as the orthodox theory for evaluating statistical hypotheses. This review serves two purposes. First of all, it rehearses the known difficulties with the orthodox position, interpreted now as a theory of inverse inference. Some suggestions are offered in Section 2.7 for mitigating these deficiencies. Second, the development of confidence interval theory from a hypothesis testing basis acts as a model for reconstructing the Fisherian theory: developing fiducial probability from significance testing.

Having motivated the analysis of Fisherian inference by rejecting Neyman-Pearson as satisfactory for inverse inference, we turn first to reconstruct Fisher's theory of significance testing. This is attempted in Chapter 3. Attention is focused on those issues which evolve into impediments for fiducial inference. The solutions offered to such problems in the theory of significance carry over to the fiducial theory.

Chapters 4 and 5 are devoted to a detailed analysis of fiducial inference. The reconstruction given in the text is similar, in a crucial respect, with the proposals made by Ian Hacking. The crux of both reformations is a novel principle of irrelevance. The need for the extra principle is demonstrated by considering a number of paradoxes which can be easily generated when relying solely on Fisher's conditions for fiducial inference. Whereas Hacking supports his principle by considerations involving *likelihood*, those presented here are defended by appeal to significance testing. Thus, the analogy with Neyman-Pearson is strengthened.

Unfortunately, it becomes clear (in Section 4.8) that a fatal shortcoming remains: one which appears even in simple problems involving only one unknown factor. The diagnosis, more accurately the postmortem, shows the disorder to be deep-seated affecting Hacking's version equally. A comparison is drawn between a well known difficulty with Laplace's Principle of Insufficient Reason and the failure of fiducial inference.

Most published criticism of fiducial probability involves multiparameter problems. Fisher spent great effort, in his last years, trying to enlarge his

theory to cope with these more complicated questions. The focus of Chapter 5 is in showing that no additional flaws need develop for fiducial inference if the precautions for the single parameter cases (Chapter 4) are observed. The step-by-step procedure formalizes the process for generalizing the single parameter solution. The reconstruction terminates with a survey of the outstanding 'paradoxes' and their resolutions based on the step-by-step technique.

Having considered, and rejected, fiducial inference as developed in Chapters 4 and 5, the discussion shifts to study two other proposals: those of Ian Hacking and Henry Kyburg (Chapters 6 and 7). Hackings's theory for fiducial inference is shown deficient as predicted in Section 4.8. Kyburg's theory is even more seriously lacking for fiducial probability, since his program violates confirmational conditionalization (CC); which is one of the conditions of adequacy set forth in Chapter 1.[6] The evaluation of Kyburg's views is based, in large measure, on results due to Isaac Levi.

Fisher's ideas on inverse inference are not limited to his fiducial probability. The investigation broadens to review suggestions made for general inverse inference: from observed samples to larger, unobserved populations. This portion of the inquiry concentrates on problems of experimental design; notably, randomization and a comparison with related aspects of fiducial theory. Here the relative merits of Hacking's and Kyburg's programs change. To accommodate the needs for random sampling, Hacking is forced to alter his stance on irrelevance (from his novel principle used in the fiducial argument), unlike Kyburg who maintains a unified position throughout.

The investigation concludes (Chapter 8) with a reconstruction and criticism of Fisher's theory of experimental design.[7] Questions surrounding principles of relevance bring to an issue the important points. The upshot is that Fisher's hopes for inverse inference (including fiducial probability at one end and randomization, seen as an essential part of experimental design, at the other end) are frustrated in two ways. Fiducial inference cannot be salvaged despite rather extreme sanctions on its application. Second, and perhaps more serious, the very natures of relevance principles intended to permit each of the two procedures clash by taking opposite stands regarding when irrelevance must be demonstrated before it can be presumed.

1.2. BAYES' THEOREM AND THE LAPLACEAN PRINCIPLE OF INSUFFICIENT REASON

It is a fundamental tenet of the Bayesian program that an ideally rational being is capable of representing his or her system of personal beliefs through

the calculus of probability. One way of formulating this requirement is to consider a language, or sets of sentences, and a function 'c(/)' defined over pairs of sentences of this language and which satisfies certain restrictions (below). This c-function is to be understood as a *confirmation* function (Carnap's logical C-functions).[8] A confirmation function captures the *rational* agent's assessment of the relative support for various hypotheses given any well formed evidential basis.

This is further explicated by way of two quasi-physchological concepts: rational credence '**Cr**' and rational credibility '**Cred**.' Let E be the total evidence, believed true and certain, of some idealized agent at some moment. Then $\mathbf{Cr}_E(h)$ gives the agent's credence for hypothesis h (on his special evidential footing at that time). Credence is a *static* (dispositional) property of the agent for a fixed epistemological state. Understand credibility as the *dynamic* generalization of this property over different evidential bases. Put simply, $\mathbf{Cr}_E(h) = \mathbf{Cred}(h/E)$. That is, credibility is the common thread holding together different credence functions.

Because of its quasi-psychological status, credence (or, for that matter, credibility too) is given an additional *temporal* component in Carnap's analysis.[9] One can at least imagine an agent changing his credibility function (hence, also his credence). If the temporal component is extracted from the credability function, then the idealized c-function is the result.

Before considering in detail the formal restrictions to be imposed on the confirmation function, the result of which is an *inductive logic*, it is appropriate to motivate a central rule governing changes in credence for special changes in the evidential base. Let $\mathbf{Cr}_E(h/d)$ be the conditional credence for hypothesis h, given the evidence E, on the *assumption* of d. Without changing the underlying credibility, it is argued that the revised credence for h on the new evidential base, call it E', formed by adding (or accepting) d and provided nothing entailed by E is deleted, is just the original conditional credence. In other words,

$$\mathbf{Cr}_{E'}(h) = \mathbf{Cr}_E(h/d).$$

That the underlying credibility is not altered throughout is explicated by the additional rule that:

$$\mathbf{Cr}_E(h/d) = \mathbf{Cred}(h/E \,\&\, d).$$

Together, the two regulations are captured by the principle of confirmational conditionalization (given below).[10]

(1) The first principle of our inductive logic imposes constraints on the c-function. The condition that the c-function be represented through the probability calculus, known as the *coherence* condition, is set forth as follows: If E is a consistent evidential basis (sentence) and h a hypothesis of the language, then

(i) $0 \leqslant c(h/E) \leqslant 1$. $c(h/E) = 1$ if h follows deductively from E.

(ii) If the hypotheses $h_1, \ldots, h_n$ are pair-wise inconsistent given E, then $c([h_1 \text{ or } \ldots \text{or } h_n]/E) = \Sigma_{i=1}^{n} c(h_i/E)$.

(iii) If e is any other evidential statement consistent with E, then $c(h/E\,\&\,e) \cdot c(e/E) = c((e\,\&\,H)/E)$.[11]

(2) A second basic condition of the Bayesian program (and from which the name is servied) is the regulation governing the change in confirmation for different (but logically related) evidential bases. Just as the agent might consider the c-function defined on the consistent set E, it is possible to evaluate confirmations with respect to the consistent basis E', defined as $E' = E\,\&\,e$. Accordingly, the Bayesians insist that the change obey Bayes' Theorem, which states that:

$$p(h/E\,\&\,e) = p(e/E\,\&\,h) \cdot p(h/E) \div p(e/E)$$

and

$$\propto p(e/E\,\&\,h) \cdot p(h/E)$$

Alternatively, the change in confirmation for the hypothesis h, reflecting the expansion of the evidential basis from E to E', is given by the rule Confirmational Conditionalization (in a form due to I. Levi).

[CC] If E and E' represent a pair of mutually consistent sentences and $e\,\&\,E = E'$, then if the conditional probability for h given $E\,\&\,e$ equals r', it follows that $p(h/E') = r'$.[12]

The justification offered by Bayesians for the coherence and CC principles brings into play yet another group of concerns namely, how beliefs (represented as probabilities) are to be used for resolving uncertainty, i.e. decision making. Assuming a well defined 'gambling' situation, including known utility-function ratings for all 'payoffs,' the Bayesians argue that by appeal to a non-controversial decision rule, i.e. maximizing expected utility, the agent can defend against certain loss (Dutch-book) by obeying coherence;[13] and by extending the gambling to 'conditional bets,' can guard against the same problem by satisfying both coherence and CC.[14] Both betting arguments are convincing for the limited domain of gambling situations meeting the numerous conditions presupposed. However, the dependence on the Dutch-book defense as a sufficient justification for the several Bayesian

requirements is hardly satisfactory, since the inclusion of scientific investigations under the caption 'bettable' is questionable, at best.

The attitude adopted here is that (1) coherence, and (2) CC represent primitive requirements on any inductive logic intended to explicate confirmation by way of mathematical probability. This does not diminish the force of the Dutch-book argument for those gambling contexts where it applies. However, when it is remembered that among the hypotheses purported to be treated by inductive logic there are included scientific laws, subsuming an infinitude of instances and extending over time unbounded by any betting conditions for payoffs, it seems unreasonable to depend on the Bayesian decision making consequences of incoherent or non-conditionalized confirmations.

(3) A third condition assumed by all inductivists, yet often left unstated or underrated for its importance, is the Total Evidence Requirement. Whereas coherence and CC serve to restrict the possible confirmation functions used to represent a body of beliefs, the Total Evidence Requirement stipulates that, at any time, the *actual* confirmation function selected by the rational agent must be one which uses *all* the evidence available at that time, and not an arbitrary subset of it. As will be shown in the sequel to this section, this restriction can be weakened to demand the inclusion of all *relevant* evidence, in place of the more extensive total evidence available. This modified principle is known as the Total Relevant Evidence [TRE] requirement.

The most serious abuses of inductive arguments result from the violation of the TRE requirement. Unlike deductive inference, the availability of additional premises (interpreted as extra evidence) can defeat a proposed inductive inference. Thus, conclusions of inductive arguments are most sensitive to modifications in the evidence available.

As hinted in the introductory section, the focus of this work falls within the domain of statistical or 'stochastic' contexts. Apart from the confirmatory or inductive interpretation of the probability calculus ('probability$_1$' in Carnap's jargon), there exist a collection of phenomena, loosely described as 'stochastic' processes resulting in 'chance' events, which supply yet another model for probability ('probability$_2$' in Carnap's jargon). Those who require, either in addition to, or in place of a probability$_1$ interpretation, an interpretation of probability which makes it an empirical property of the real world are called Frequentists. The common bond unifying the Frequentists is the commitment to some long-run properties of stochastic processes, loosely described as limiting frequency properties of idealized repeated trials of the stochastic mechanism which bears the probability$_2$ property.

For example, it is supposed that the frequencies of 'heads' and 'tails' in repeated flips of a 'fair' coin will (eventually) approximate the 50%–50% frequency count. However no one has given a satisfactory reduction of the probability$_2$ concept to either actual frequencies, or even to hypothetical extensions of actual frequencies, which does not presume greater metaphysical truths than the Frequentists (on the whole) are willing to grant. This includes both von Mises'[15] and Reichenbach's[16] interpretations, if their infinite sequences and mathematical limits of relative frequencies are intended to supply the truth conditions for the corresponding probability$_2$ statements. They fail, also, to provide verification conditions based on the available (finite) evidence, or else they presuppose other probability$_2$ statements of even greater metaphysical consequence.

No proposal is made here for a satisfactory explication of the probability$_2$ concept. Nonetheless, the Frequentists' interpretation of probability is allowed, at least so far as to permit the following Frequency Principle (due, in spirit, to Ian Hacking):[17]

(4) If the probability$_2$ for an event E is r and e is a statement to that effect, then the confirmation (probability$_1$) for the hypothesis h, that E occurs, given the evidence e, is just r. Alternatively,

$$\mathrm{p}(h/e) = r.$$

The Frequency Principle provides the means for using knowledge of stochastic probability to determine confirmational probability. Among all but the most radical Bayesians who refuse to admit stochastic probability, e.g. deFinetti, Frequentism is accepted to the degree that the Frequency Principle gives a satisfactory resolution to 'direct inference.' Thus, when flipping a 'fair' coin or picking a card from a 'well shuffled' deck, or in any one of a number of other situations with known stochastic properties, it is hardly debatable what the 'probability' is, or ought to be, that an arbitrary ('random') outcome of the stochastic process is such and such. Knowing that the coin is 'fair' determines the probability that an arbitrary flip results in 'heads.' It is .5, by the Frequency Principle. This does not mean that all participants would go as far as to adopt a confirmational interpretation of the individual 'probability' concept. But the important point is that such an interpretation could be applied, and it would satisfy the desired principles (1)–(4).

The almost universal agreement for cases of direct inference suggests the following additional condition, which (for lack of a better name) is called the *objectivity* rule:

(5) There exists exactly one c-function which represents the confirmation $\mathrm{p}(h/BK)$; where h is some hypothesis and BK is the total (relevant) background evidence available. The title 'objectivity' refers to the observer *invariant* status of the confirmation function, given a fixed evidential basis.

Admitting, then, that *objectivity* presents little hindrance to a Bayesian theory of inductive logic when the evidential basis is relatively 'rich,' so that the Frequency Principle can be employed, the challenge is to develop Objective Bayesianism to cover cases of minimal information; typified by 'inverse inference' problems. In order to see that some *additional* principle is needed for this end, it is appropriate to review the classic solution and its Frequentist objections.

The oldest resolution of this problem, due mostly to Laplace[18] (although seen by many as a systemization of a suggestion credited to Bayes), is found in the so-called Principle of Insufficient Reason (also known as the Principle of Indifference). It states that where the background knowledge is strong enough merely to list a set of 'possibilities' for some unknown factor, then the correct confirmation ('probability'), given just this basis, is one which assigns equal values to the various possibilities. In conjunction with Bayes' Theorem and CC it provides a powerful solution to the inverse inference problem.

For example, suppose that a 'chance' event, the flip of a coin, can be modelled by the mathematical distribution known as the 'binominal' distribution. That is, the 'bias' of the coin is represented by a parameter θ whose values are restricted to the interval [0, 1] and such that given the knowledge that $\theta = k$ (k a known constant), the probability$_2$ is k that a flip will fall 'heads-up.' In addition, it is assumed that the repeated flips are stochastically independent, so that the probability$_2$ for any sequence of flips (represented by a conjunctive proposition in the probability$_1$ formulation) is the product of the probabilities for the simple events (represented by the simple conjuncts). It follows that, in general, the probability$_2$ of getting r 'heads' out of n such trials, given the information that $\theta = k$ is:

$$C_n^r \cdot k^r \cdot (1-k)^{n-r};$$

where C_n^r is the binominal coefficient equal to $n!/r! \cdot (n-r)!$. This probability$_2$ value supplies the answer to the 'direct inference' question, for which there is little controversy based on the Frequency Principle.

However, the interesting cases arises when BK does not contain the information about the value of θ, and on the basis of sample data of previous flips it is desired to assess the confirmation for hypotheses about other, still

unobserved, chance events. Letting the evidential record of the past n flips of the coin be represented by the conjunctive proposition e, it follows from Bayes' Theorem and CC that:

$$\mathrm{p}(\theta = .5/BK \,\&\, e) = \mathrm{p}(e/BK \,\&\, (\theta = .5)) \cdot \mathrm{p}(\theta = .5/BK) \div \mathrm{p}(e/BK). \quad [1]$$

Using Axioms (ii) and (iii), it follows that:

$$\mathrm{p}(e/BK) = \int_{\theta} \mathrm{p}(e/BK \,\&\, (\theta = \theta^*)) \cdot \mathrm{p}((\theta = \theta^*)/BK) \cdot d\theta. \quad [2]$$

By substituting [2] in the denominator of [1], a more familiar form of Bayes' Theorem is reached.[19]

Therefore, relying on the direct inference solution to provide values for: $\mathrm{p}(e/BK \,\&\, (\theta = \theta^*))$, for which $\mathrm{p}(e/BK \,\&\, (\theta = .5))$ is but a special case, the Objective Bayesian answer to the inverse inference question is completed by using the Principle of Insufficient Reason to provide the confirmation: $\mathrm{p}((\theta = \theta^*)/BK)$. Since the parameter θ can take any value in the interval [0, 1], the Indifference Principle recommends (in its simplest form) the initial or prior confirmation which is uniform over the entire permissible range. Hence, the probability (density) is: $\mathrm{p}((\theta = \theta^*)/BK) = 1$ for $\theta^* \epsilon$ [0, 1], and 0 otherwise. This 'prior' probability (density) selection generates a 'posterior' probability (density) for $\mathrm{p}((\theta = \theta^*)/BK \,\&\, e)$ (for which $\mathrm{p}(\theta = .5/BK \,\&\, e)$ is but a special case) in agreement with the ever popular solution of T. Bayes', dating from 1763.[20] Continuing with this line of reasoning, we see that the calculus of probabilities can be used to generate a simple quantitative law for future (or other unobserved) chance events, covered by the same binomial model. The well-known Rule of Succession gives the law exactly.[21] It states that the probability for the next chance event showing property F (given that r of the previous n trials had the property) is $(r + 1)/(n + 2)$.

Serious criticism of the Indifference Principle as an adequate solution for the assignment of initial or prior probabilities did not take hold until the publications of G. Boole (1854)[22] and J. Venn (1866):[23] they encouraged the demand for a 'frequency' interpretation of the probability or confirmation function, as well as the 'inductive' or 'logical' interpretation. Before considering this 'defficiency' of the Indifference Principle, it is best to review a more telling charge, internal to the objective Bayesian program: that it succumbs to inconsistency.

As stated above, the principle assumes a well defined set of epistemic alternatives, which receive equal prior support. Unfortunately, different

formulations for the same alternatives can lead to mutually inconsistent assignments. The most striking cases arise when the alternatives can be depicted, mathematically, by a continuum of values. Adopting a uniform initial confirmation under one representation of these values is incompatible with a uniform initial confirmation when the values are transformed (1-1, or point by point) in a non-linear manner.[24]

A representative version of this 'paradox' (repeated on p. 132) is as follows. Suppose that a liter of a wine-water mixture is known only to contain between 50% and 90% wine. Therefore, the ratio of wine to water varies from 1:1 to a maximum of 9:1. Applying the Principle of Insufficient Reason, the a priori probability for the contents of the mixture is uniform over the various possibilities, which form a continuum of values. However, expressed in terms of percentage wine, the prior probability is $\frac{1}{2}$ that there is more than 70% wine; whereas, expressed in terms of the ratio of wine to water, the prior probability is $\frac{1}{2}$ that the ratio is greater than 5:1. But this last value corresponds to a minimum of at least (approximately) 83.5% wine – which is inconsistent with results in terms of percentage wine. Unrestrained application of the Principle of Indifference leads to incoherent results.

Among contempory Objective Bayesian programs which rely on a modified version of the Laplacean Indifference Principle, Harold Jeffreys' is unquestionably the most sophisticated, reflecting sensitivity to both intricate mathematical problems and delicate philosophic issues.[25] He avoids the foregoing pitfalls by invoking subtle invariance conditions to guarantee consistency, ahead of time, over all potentially acceptable alternative mathematical representations for the unknown parameter(s). However, his solutions remain unconvincing to the majority of readers, primarily because the conditions of adequacy for evaluating invariance are unintuitive and sufficiently vague to preclude unique answers for all cases. It is to Sir Harold's credit, however, that in the forty-five years since his work was first published no one has managed to produce a more satisfactory formulation of the invariance or symmetry conditions which is both moderately plausible and general enough to permit practical application.[26]

In summary, the aforementioned criticism of the Indifference Principle can be seen as a charge of *incoherence* of the resulting confirmational or probability$_1$ theory. Of course, those who insist on retaining a modified version of the principle find it necessary to restrict the application of the rule; hence, achieving coherence at the cost of generality. It has been argued that the mode of restricting the principle, through various symmetry or invariance conditions on the mathematical representation of the unknown

factor(s), proves inadequate because it fails to provide unique solutions for all cases (*objectivity*) and because of the ad hoc nature of the defining invariance regulations.

Tracing the historical debate over the Laplacean principle, the objections raised by Boole and Venn are of a different nature and antedate the incoherence problem, discussed above. The version of Frequentism which they advocated requires that the Frequency Principle account for *all* confirmational relations between data and hypothesis. Their strict Frequentism has, as an underlying motivation, the attitude that only the probability$_2$ concept is a legitimate explication for 'probability' and the confirmational interpretation has validity indirectly; solely by inheritance from the probability$_2$ evidence and through the Frequency Principle. Thus, all confirmational statements presuppose an empirical fact; to wit, the probability$_2$ property true of the event in question in the hypothesis.

Therefore, for Venn and Boole, applying the Indifference Principle to allow the probability$_1$ claim that, for example, the a priori probability for the 'bias' of the coin is uniform over the [0, 1] range, presupposes (without warrant) an empirical fact about the *stochastic* genesis of the coin. That is, in accord with the strict Frequentist doctrines, the a priori values from the Indifference Principle could be used if the coin were selected, at random, from an urn containing variously biased coins and such that the proportion of coins of equal 'bias' were uniform for all biases. In other words, if the Frequency Principle in conjunction with the probability calculus could be used to generate the prior probability$_1$ values, then there would be no objection from the strict Frequentists to the use of the Laplacean rule.

However, it is just for those cases where the evidential basis is insufficient for the determination of initial or prior support via the Frequency Principle that the Laplacean principle applies. Of course, if the demands of the strict Frequentists could be met in all cases, then there would be little difference between problems of 'direct' and 'inverse' inference. This is because all the confirmational calculations, usually reserved for inverse inference and utilizing Bayes' Theorem, could be carried out within the probability$_2$ concept, so that only the Frequency Principle would be needed for calculating probability$_1$.[27]

To summarize the criticisms voiced by Boole and Venn, not only does the Frequency Principle apply for 'direct inference' but it comprises a *complete* inductive logic (when conjoined with (1) coherence and (2) CC).[28] If the evidential basis is too weak to permit the application of the Frequency Principle, the usual strict Frequentist response is to deny any confirmational

relation capable of supporting a probabilistic interpretation. In these cases a hierarchy of secondary and tertiary methods are employed to mitigate the absence of full-fledged 'inverse inference.' For illustration, in the Neyman-Pearson theory (to be discussed in the next section) the policies of *hypothesis testing* and *confidence intervals* fill the gap left by 'inverse inference.'

In Fisher's theory (discussed in great detail in Chapters 3, 4 and 5) *significance testing, likelihood ratio support* and *fiducial probability* supply the missing components of his Frequentist program.[29] The position to be attributed to Fisher is a weakened version of Frequentism, i.e. which incorporates the Frequency Principle, but which allows for additional rules of inductive inference. The essential features of this modified Frequentism are: (*a*) the adoption of a two-fold interpretation of probability – using both probability$_1$ and probability$_2$; (*b*) the commitment to a confirmational theory which satisfies: (1) coherence, (2) CC, (3) TRE principle, (4) the Frequency Principle, (5) objectivity, and finally, the use of only 'frequency based' rules to complete the theory.

Thus, in place of the Laplacean Principle of Indifference, or modifications appealing to symmetries formulated in terms of the representation of the unknown parameter(s), the Fisherian theory is completed by adopting principles stated only in terms of the probability$_2$ evidence already available. These new rules surpass the inductive powers of the simple Frequency Principle, allowing for confirmation in cases of inverse inference, i.e. where no prior support is available using just the Frequency Principle. The justification for these additional principles, apart from the added scope given the new theory, rests on the interweaving of older Frequentist procedures, e.g. significance testing, with a general Objective Bayesian program. Thus, the position attributed to Fisher is seen to fall between strict Frequentism (at the one extreme) and Objective Bayesianism based on the Laplacean Principle of Insufficient Reason (at the other).

Before embarking on a detailed reconstruction of Fisher's program, it is to the point to review briefly the received strict Frequentist position, as it is presented in the Neyman-Pearson theory of hypothesis testing and confidence intervals. This program does not attempt to duplicate the confirmational properties of the Objective Bayesian position, but by relying on the Frequency Principle and certain pragmatic factors to alleviate the TRE requirement it parodies the Bayesian answers to inverse inference through the two concepts of hypothesis testing and confidence intervals. To emphasise the theoretical difference between their program and the Bayesian one, Neyman-Pearson relied on the ordinary language meaning of such terms

as *behavior* (in the technical term 'inductive behavior'); as opposed to *inference*, and the characterization of their program as fundamentally 'decision theoretic.'

Unfortunately, despite this careful choice of language, the usual attitude taken towards the Neyman-Pearson program (at least by English speaking statisticians) is as a proposed theory of inductive logic. In the chapter that follows, the theory will be evaluated to demonstrate, once again, its inadequacy for this (unintended) purpose. The perspective from which the criticism is made has been most elegently supplied by Ian Hacking in his text *Logic of Statistical Inference.*[30] To give preliminary formulation of that view now, it is argued that the N-P theory is plausible as a theory of inference *before seeing the actual evidence* (on the 'forward' look), but fails as a theory of inference *after seeing the data* (on the 'backward' look). Of course, it is the goal of the reconstructed Fisherian theory to satisfy both.

The attention to the N-P theory of confidence intervals, and its careful reconstruction, pays dividends in excess of just understanding the strict Frequentist proposals for answering the inverse inference questions. Both Fisher's Fiducial probability and Neyman's Confidence Interval probabilities were developed within a five year span of each other: Fiducial probability in 1930 and Confidence Intervals in 1934. Because of the exact *numerical* agreement between these two methods for a host of common statistical problems it was thought by both involved parties that the two concepts were equivalent: at least the agreement was kept up publicly through 1935.[31] However, like the now infamous Behrens-Fisher problem (discussed in detail on pp. 63 and 146) cases were found to establish the numerical independence of the two procedures, i.e. confidence interval solutions which failed as fiducial probabilities and vice versa.

To give the reader a taste for the intimate connection between the two methods, and as a preliminary example of the informal arguments used to derive the several solutions, consider the following heuristic illustration. Suppose that a 'chance' event is to be observed, and the background evidence includes a single unknown parameter, denoted θ; with a set of epistemically possible values, denoted Δ. For each value of $\theta \epsilon \Delta$ the distribution of the observable, denoted x (and called a random variable) is known. For a given value of θ, let m_θ designate the *median* value of X, i.e. the value of X such that the probability is $\frac{1}{2}$ of getting a value less than (or greater than) m_θ.[32] Hence, before seeing x and by use of the Frequency Principle, the probability$_1$ for the hypothesis: $x \geqslant m_\theta$, given the background described above, is exactly .5 (without any further restriction on θ). By way of emphasizing the impor-

tance of this inference, it would retain its inductive validity even if the evidential basis were augmented with frequency based prior probabilities. In other words, the probability$_1$ statement (above) is independent of the prior probability for θ.

Using the N-P formulation of the counterpart solution to the inverse inference problem for 'probability' about θ, given knowledge of x, one can retain the probability$_1$ statement even after x becomes available (for calculational purposes). That is, knowing $x = 8$ (for instance), permits us to conclude that with confidence .5, $8 \geqslant m_\theta$.

This induces a (confidence) interval for θ. The confidence interval (at the .5 level) for hypotheses about the unknown parameter, given observation of the random variable, is *not* to be construed as probability$_1$ for θ given BK and $x = 8$, but as probability$_1$ for θ, given BK; where '$x = 8$' is used for computational purposes only. Exactly what 'computational purposes' means, and how it is to be distinguished from using the data as evidence, i.e. including it in the background evidence available, is the focus of the difference between the 'forward' and 'backward' looks.

The Fisherian argument for deriving the fiducial probability in this problem is similar, only the probability$_1$ conclusion of the direct inference (above) is supposedly retained given BK and '$x = 8$' as new evidence. Exactly how the observation can be included in the evidential basis whilst retaining the conclusion of the direct inference is the key to the reconstruction of fiducial probability.

This chapter concludes with a review of several standard Bayesian concepts, pertinent to the notion of inductive relevance, and also with a simplified presentation of the extended probability calculus; to allow for continuity – both in the parameters used to represent the unknown factors of inverse inference, and in the random variables used to represent the empirical data or observations.

1.3. INDUCTIVE RELEVANCE AND RELATED CONCEPTS: LIKELIHOOD, SUFFICIENCY AND ANCILLARITY

This somewhat condensed section is intended to remind the reader of several basic results concerning inductive relevance. Throughout, it is assumed that (1) coherence and (2) CC apply, so the discussion has pertinence for the Bayesians as well as the more restrictive programs, e.g. strict Frequentists.

Intuitively, evidence is *relevent* to a hypothesis if it plays some role in the inductive assessment of the hypothesis. In other words, the probability

assigned the hypothesis depends on the availability of *relevent* evidence. This idea is formalized in a definition, due to Carnap.[33]

DEFINITION: Data e are *relevant* to the hypothesis h, relative to the background evidence BK, just in case:

$$p(h/BK) \neq p(h/BK \& e).$$

Using this notion the Total Evidence requirement can be weakened without altering the confirmational relation between hypotheses and evidential bases. Instead of demanding that the agent cite all the evidence currently in hand, which might include all sorts of intuitively irrelevant material, it is necessary only that the *relevant* portion be used. Hence, from the definition of relevance (above), it is demanded that the agent cite all the *relevant* evidence available when evaluating inductive support. Of course, there need not be a unique evidential sub-basis which is *the* total relevant portion. However, the modification to a Total Relevant Evidence [TRE] condition does substantially ease the burden of *applying* the theory.

It is often claimed, even by statisticians themselves, that a fundamental part of statistics is the ability to condense large quantities of numerical data into usable summaries. That is, one tries to extract from masses of quantitative material a kernel of relevant information which fully captures the evidential content of the original facts. Not only is this summary a matter of convenience, e.g. for communicating the results economically or for allowing simple calculations (neither of which are terribly important with the avalanche of computerized methods), but a proper summarization can quickly point out where experimental design is important and even suggest improvements for future inquiry.

With this goal in mind, it is desired to develop general principles of relevance which can be applied no matter what the particular evidential basis of the agent might be. That is, if one wants to publicize the outcome of some elaborate experiment, including several hundred individual results, subject to a modicum of practical constraints, them it is highly desirable to be able to compact the data into a concise summary report *without loss of relevant information.* However, the obvious problem is how to insure that what counts as a legitimate summary for one investigator, with his individual background knowledge and concerns, remains a valid report for all other interested parties?

The first assumption to be made in advance of a general solution is that the family of investigators who are to make use of the public report share a common inquiry. That is, it is to be required that the summary be adequate

only with respect to a clearly specified set of alternative hypotheses, if not a single hypothesis. Loosely speaking, where the problem is 'stochastic,' a shared agreement, expressed as the acceptance of a common 'model,' serves to delimit the range of hypotheses to be considered, This presumption carries with it another benefit. By assuming a common 'model' for all concerned parties it is guaranteed that they also share a common set of beliefs; at least so far as these relate to the phenonema under investigation. An artificial, but otherwise satisfactory example, is as follows. Investigator Jones is uncertain about the 'bias' of a newly minted coin (which henceforth is to serve as the standard 'flipping' coin for resolving all sorts of decision problems). He has experimented with the coin by flipping it (under regular conditions) several hundred times and recorded the sequence of outcomes. Now, he wishes to publish his analysis including the experimental data.

The common model for this case is the 'binominal' distribution. That is, it is assumed by Jones and all other investigators concerned that the 'bias' of the coin can be represented by the binominal parameter θ, restricted to the [0, 1] interval, and that the repeated flips are stochastically independent. This last assumption is spelled out as insisting that the probability$_2$ for any n-fold sequence of flips be equal to the *product* of the n probability values, no matter what the order of the n-fold sequence. In general, two events are stochastically independent for a given parameterization iff

$$\mathrm{p}(e_1 \,\&\, e_2/\theta) = \mathrm{p}(e_1/\theta) \cdot \mathrm{p}(e_2/\theta);$$

where the probability function is interpreted as probability$_2$. By the direct inference, based on the Frequency Principle, stochastic independence transfers to confirmational or probability$_1$ independence.

With all these suppositions accepted, Jones' problem becomes one of constructing a report of his experiment, short of detailing all the outcomes and in their exact sequence, which captures all the relevant aspects of the data, considered as evidence. Note that by adopting a 'model' Jones has not assumed that he is in perfect agreement with the other investigators. There is still a lot of room for personal differences. reflected in the adoption of *different* prior probability$_1$ assignments for the values of θ. Therefore, Jones is looking for some summary of his data which retains all the relevant evidence about θ, *no matter what the prior probability for* θ.

The solution to this problem is found by reviewing the CC principle, in conjunction with Bayes' theorem. Let e represent Jones' data in its entirety. Subject to the various assumptions (above), it follows that the posterior probability$_1$ for hypotheses about θ is given by:

$$p(\theta/e\,\&\,BK) = p(e/\theta\,\&\,BK)\cdot p(\theta/BK)/p(e/BK)$$

$$= p(e/\theta\,\&\,BK)\cdot p(\theta/BK)/\int_{\theta} p(e/\theta\,\&\,BK)$$

$$\cdot p(\theta/BK)\cdot d\theta,$$

and

$$\propto p(e/\theta\,\&\,BK)\cdot p(\theta/BK);$$

where $p(e/\theta\,\&\,BK)$ is given by the direct inference, i.e. via the Frequency Principle; $p(\theta/BK)$ is the prior probability for θ; and $p(e/BK)$ is the normalizing constant.

A concept due to R.A. Fisher,[34] and dating from about 1912, is the *likelihood function;* defined as:

$$p(e/\theta\,\&\,BK)$$

interpreted as a function of the parameter θ (not as a function of the data e). Likelihood, therefore, is what is left of Bayes' Theorem after the problematic prior probability is deleted. Fisher, as more recently G. Barnard and A.W.F. Edwards, advocates the use of the likelihood function as a measure of inductive support. Generally, they recommend that the support for various hypotheses be assigned in direct proportion to the likelihood for the hypotheses, given the data e. Now it is clear that this solution to 'inverse inference' will be consistent with the Bayesian solution so long as the prior probability is uniform for all values of θ. However, what is more important for the question at hand, the replacement of e by the likelihood function for e permits the Bayesian to generate his own posterior probability by inserting his own prior probability factor, in accord with Bayes' Theorem. Therefore, one answer to Jones' question is the use of the likelihood function as a summary of the data e. In many cases it is much easier to report.

In accord with common usage, define a summary (like the likelihood function) which preserves all the relevant evidence, no matter what the prior probability, to be a *sufficient* summary of the original data. Next, define a real-values function of the data to be a *statistic* of that data. Finally, the sort of summary required by Jones is one which is a *sufficient statistic* of his data. For a one parameter model, a more formal definition of a sufficient statistic is available.[35]

DEFINITION: Let $t(x)$ be a statistic, i.e. a real-values function, of the data x. Then $t(x)$ is a sufficient statistic (of X) for the parameter θ iff (suppressing BK):

$$p(x/\theta) = p(x/t(x)) \cdot p(t(x)/\theta);$$

where $p(x/t(x))$ is independent of the unknown parameter.

That a sufficient statistic, so defined, accomplishes the task of summarizing the original data x without loss of relevant evidence, is demonstrated by the following argument.

From Bayes' Theorem and CC the posterior probability for hypotheses about θ is given as:

$$p(\theta/x \,\&\, BK) \propto p(x/\theta \,\&\, BK) \cdot p(\theta/BK) =$$
$$p(x/t(x) \,\&\, BK) \cdot p(t(x)/\theta \,\&\, BK) \cdot p(\theta/BK)$$

and

$$\propto p(t(x)/\theta \,\&\, BK) \cdot p(\theta/BK)$$

since $p(x/t(x)\&BK)$ is, by assumption, independent of θ and can be absorbed into the constant of proportionality. Therefore, by reporting just the sufficient statistic $t(x)$, the relevant portion of the data x is saved, while discarding the rest; which is demonstrably irrelevant to θ.

For illustration, returning to Jones' problem, from the binominal distribution for independent trials, the probability for the outcome e, given θ and BK, equals $\theta^r \cdot (1-\theta)^{n-r}$; where r out of n flips fell 'heads-up' and θ is the 'bias' of the coin for falling 'heads-up.' If $t(e)$ is the proportion of flips resulting in 'heads' and C_n^r is the binomial coefficient then

$$p(t(e)/\theta \,\&\, BK) = C_n^r \cdot \theta^{t(e)\cdot n} \cdot (1-\theta)^{(1-t(e))\cdot n}.$$

Also, $p(e/t(e)\&BK) = 1/C_n^r$, which is independent of θ. Therefore, $p(e/\theta \,\&\, BK) = p(e/t(e)\&BK) \cdot p(t(e)/\theta \,\&\, BK)$ and $t(e)$ is a sufficient statistic for the unknown bias.

Another derived relevance condition, actually an *irrelevance* condition, is ancillarity.[36]

DEFINITION: A statistic $t'(x)$ is called an *ancillary* statistic (of the data x) for the parameter θ iff (suppressing BK)

$$p(x/\theta) = p(t'(x)) \cdot p(x/\theta \,\&\, t'(x));$$

where $p(t'(x))$ is independent of the unknown parameter θ. It is claimed that an ancillary statistic represents a reduction of the data which contains *no* relevant information about the unknown parameter. Appealing, again, to Bayes' Theorem and CC

$$\begin{aligned} p(\theta/x\,\&\,BK) &\propto p(x/\theta\,\&\,BK)\cdot p(\theta/BK) = \\ &\quad p(t'(x)/BK)\cdot p(x/t'(x)\,\&\,\theta\,\&\,BK)\cdot p(\theta/BK) \\ &\propto p(x/t'(x)\,\&\,\theta\,\&\,BK)\cdot p(\theta/BK) \\ &\propto p(\theta/x\,\&\,BK\,\&\,t'(x)). \end{aligned}$$

Therefore, by conditionalizing on $t'(x)$ first, i.e. by augmenting the background evidence BK to include $t'(x)$, there is no change in the prior probability for θ, and the posterior probabilities (given BK & $t'(x)$ & x or BK & x) are equivalent.

The importance of ancillarity is not at once evident. However, it does play a critical part in the evaluation of certain well known statistical procedures, e.g. randomization. More to the point of this inquiry, the use of ancillary statistics to generate conditionally sufficient statistics (known as *exhaustive* statistics) involves the fiducial argument (see pp. 112–116).

1.4. CONTINUITY: THE CUMULATIVE DISTRIBUTION FUNCTION AND DENSITY FUNCTION

Besides the application of the probability calculus to stochastic processes with discrete outcomes, e.g. the binomial distribution applied to flips of a coin, it is possible to extend the interpretation to cover continuous cases as well. The assumption that chance events can be continuously variable represents mathematical convenience, without literal veracity. Real observations are discrete, of course, due to the limited accuracy of instruments used.

Presupposing continuity as an approximation, probability$_2$ is obtained from a *cumulative distribution function* [c.d.f.]. Let the chance outcome be represented by a continuous random variable x. The c.d.f., denoted $F(x)$ (or $F(x, \theta)$ when a single free parameter is included in the stochastic model), assigns to any real number $x_0 (= x)$ the probability for the set of possible values of X less than or equal to x_0. The portion of the domain of F corresponding to possible outcomes of X is called the (positive) sample space. Clearly, F is non-decreasing for non-decreasing values of x.[37]

The c.d.f. for discrete random variables, e.g. binomial frequency counts, is a discontinuous 'step function.' For continuous random variables it is usually assumed not only continuous, but also differentiable. That is,

$$F(x) = \int_{-\infty}^{x} f(y)dy \text{ and } f(x) = \partial F(x, \theta)/\partial x.$$

Here f is the *density function* for the c.d.f. F and random variable x It is essential to note that a density function depends not only on the value of free parameters (if there are any), but also the precise mathematical form of the random variable. Transforming the random variable x to x' changes the density function for x'. That is, the density function for x is proportional to $F(x+\delta)-F(x)$, as $\delta \to 0$; the constant of proportionality being $1/\delta$. However, when the tranformation is employed to switch random variables (call it T), the resulting density function for x' need not be simply related since the constant of proportionality is not preserved by T. That is, it is assumed that T is measure preserving, so that $F(T(x+\delta))-F(T(x))=F(x+\delta)-F(x)$. But $T(x+\delta)-T(x)$ need not be proportional to δ. For example, if $T{:}x \to x^3$, then the resulting density for $x'(=x^3)$ is related to the density function for x by the function $3x^2$.

Finally, note that densities, apart from being finite and non-negative, are not bounded above as are probabilities. It is to be remembered that, strictly, the probability that $x=x_0$ is always 0 (for the continuous cases considered). Loosely speaking, the density function gives the relative probability for an observation 'in the neighborhood' of a specified value of the random variable.

Multivariate continuous distributions are defined in a similar manner: (for simplicity, all parameters are held fixed)

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv \text{ and } f(x,y)=\partial^2 F(x,y)/\partial x\,\partial y.$$

The connection to the c.d.f. for each of the two variables (x and y), call them $F_1(x)$ and $F_2(y)$, is also clear. For example,

$$F_1(x)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(u,v)\,du\,dv,$$

and is known as the *marginal distribution* for x. Also, there is a *marginal density* for x, given as

$$f_1(x)=\int_{-\infty}^{+\infty} f(x,v)\,dv.$$

Most importantly, if the marginal density for an observation $x=x_0$ is not zero, then one can define a *conditional density*, given x_0 as:

$$f_2(y/x_0)=f(x_0,y)/f_1(x_0).$$

Conditional densities obey an analogous generalization of Bayes' Theorem; to wit:

$$\begin{aligned} \mathfrak{f}_2(y/x_0) &= \mathfrak{f}_1(x_0/y) \cdot \mathfrak{f}_2(y)/\mathfrak{f}_1(x_0) \\ &\propto \mathfrak{f}_1(x_0/y) \cdot \mathfrak{f}_2(y). \end{aligned}$$

Therefore, all the special interpretations of Bayes' Theorem for the discrete case, e.g. in relation to CC, are extended to the continuous case by using conditional densities. Of course, this presumes an extension of the Frequency Principle to cover direct inference with continuous random variables.

Besides considering continuous random variables, it is perfectly natural to include continuous parameters within the scope of a confirmation function. Where the background information fails to provide a confirmational (or probability$_1$) density function for the parameter by way of a direct inference, a generalized inverse inference problem results. The Laplacean principle (see p. 10) then provides for a prior density over the unknown parameter.

Where the parameter is bounded both above and below, the use of a 'uniform' ignorance prior density is well defined once the mathematical form of the parameter is fixed. However, with unbounded parameters, e.g. $\theta \in (-\infty, +\infty)$, the resulting density in *improper*, since $\int_{-\infty}^{+\infty} d\theta$ is unbounded. Without reviewing the many arguments provided by Bayesians, it is generally felt that the 'correct' ignorance prior probability for such cases is improper. Perhaps the most compelling interpretation for the improper density is provided by an argument due to A. Rényi.[38]

The improper density function is not to be interpreted literally, but is a limiting representation for ignorance. That is, if it became known (only) that the parameter θ were covered by some bounded interval $[a, b]$, then the claim of ignorance, beyond this fact, for statements about θ would, using Insufficient Reason, assign equal probability to all values in the interval; hence a uniform distribution over $[a, b]$. Next, let the length of the interval grow uniformly. What results, in the limit, is the improper uniform distribution. Also, by holding the improper 'ignorance' prior, subject to this interpretation, one is committed to a policy of assigning proper, uniform distributions to the interval $[a, b]$ conditionally: on the additional information that $\theta \in [a, b]$.

This concludes the brief review of the extension of the probability calculus to cover continuous random variables and parameters. The discussion here has been only superficial, especially with respect to assumptions of differentiability and the use of improper density functions. Differentiability problems will not be of much importance for the subject of this investigation.

A more thorough treatment of the standard difficulties is available in W. Feller's introductory text.[39] The use of improper densities is, however, not unproblematic. As will be hastily considered (see pp. 161–164), multivariate fiducial probabilities are not immune to contradictions arising from a casual treatment of the improper densities. Some pertinent remarks on this subject are reserved for that phase of the examination.

Subject to some important results due to D.V. Lindley (discussed later on), the fiducial argument can be made coherent only for certain types of distributions (in one parameter cases). These have the common feature of being location parameters (or transformable thereto). Following Lindley,[40] a *location parameter* θ is defined as one such that the distribution for the random variable x depends only on the difference $(x-\theta)$. In other words, the 'shape' of the distribution does not change as θ varies throughout the parameter space. Only the location of 'key' features, e.g. the mode or median of the distribution, alters relative to fixed points on the x-axis.

Another important distribution form is that of *scale parameter*. In a parallel fashion, θ is said to be a scale parameter if the distribution depends only on the product $x \cdot \theta^{-1}$. Scale parameters can be converted to location parameters by suitable transformations. For example, assuming both the parameter and variable positive, the density for a scale parameter depends only on $\exp[\ln x - \ln \theta]$. Absorbing the exponential component into the distribution function, switching to random variable $\ln x$ and to parameter $\ln \theta$, yields location parameterization.

NOTES

[1] See, for example, Bruno deFinetti, 'Foresight: Its logical Laws, Its Subjective Sources', trans, H.E. Kyburg, Jr., in Henry Kyburg, Jr. and Howard Smokler eds., *Studies in Subjective Probability* (New York: John Wiley & Sons, 1964), pp. 93–158.

[2] Ronald A. Fisher, 'Inverse Probability,' *Proceedings of the Cambridge Philosophical Society* xxvi (1930), 528–535. Also, rep. in *Contributions to Mathematical Statistics* (New York: John Wiley & Sons, 1950), paper #22, p. 527a (with author's introduction). Hereafter cited as *CMS*.

[3] Jerzy Neyman, 'On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection,' *J. Royal Stat. Society* 97 (1934), 558–625; especially note 1. Also rep. in *A Selection of Early Statistical Papers of J. Neyman* (California: University of California Press, 1967), paper #10, p. 98. Hereafter cited as *SESPJN*.

[4] R.A. Fisher, *Statistical Methods and Scientific Inference*, 3rd ed. (New York: Hafner Press, 1973).

[5] D.V. Lindley, 'Discussion – Conference on Fiducial Probability', *Bulletin Institut International de Statistique* 40[2] (1963), 920. Hereafter cited as *Bull. Inst. Int. Stat.*

[6] Kyburg rejects CC in *The Logical Foundations of Statistical Inference* (Dordrecht and Boston: D. Reidel, 1974), Chapter 11. Objections raised here (Chapter 7) question his justification for violating CC.

[7] Only a limited range of experimental situations are treated. See note 3 of Chapter 8 for details.

[8] Rudolf Carnap, 'Inductive Logic and Rational Decisions,' In Carnap and Richard Jeffrey eds., *Studies in Inductive Logic and Probability* (California: University of California Press, 1971), p. 25.

[9] *Ibid.*, p. 14.

[10] This version of the principle is taken from I. Levi, 'On Intermediate Probabilities', *Journal of Philosophy* LXXI, #13 (1974), p. 397. It is a variant of requirement R3 found in Carnap, p. 15.

[11] Axioms (i)–(iii) are based on the Renyi system [Alfréd Rényi, 'On a New Axiomatic Theory of Probability'. *Acta Math. Acad. Sci. Hungaricae* 6, 285–335. But beware of his theorem 11. The text is based on the formulation given in D.V. Lindley, *Introduction to Probability and Statistics* (Cambridge: Cambridge University Press, 1969), p. 6] which differs from the more usual Kolmogorov axiomatization [A.N. Kolmogorov, *Foundations of the Theory of Probability*, (New York: Chelsea, 1956), pp. 2–3] in that conditional probability 'p(/)' is taken as primitive in the former; whereas absolute probability 'p( )' is primitive in the latter (and conditional probability is defined). Also the usual formulation of the probability calculus is in terms of sets, so that probability is a measure on a field of sets. The foregoing three conditions represent the translation from the typical 'measure on a set of events' language to 'propositional' language, permitting the c-function interpretation of the calculus. Finally, condition (iii) is stated only for relatively consistent evidential bases in keeping with the initial restriction to consistent base E. The important, but very complex question of how to handle inconsistent evidential bases is bypassed by assuming that ideally rational beings do not succumb to 'errors' as gross as logical inconsistency. This, of course, does not excuse a fully practical theory from the need of treating such problem cases. It is beyond the scope of the endeavor here to engage in such enquiries.

[12] In the remarks (above) and henceforth (unless so stated) the c-function will be replaced by a 'p(/)' or probability function; thereby collapsing the distinction between inductive support and probability *interpreted as confirmation.*

[13] For example, see, John Kemeny, 'Fair Bets and Inductive Probabilities', *Journal of Symbolic Logic* 20, #3 (1955), 263–273. Also, R. Sherman Lehman, 'On Confirmation and Rational Betting,' pp. 251–262 of the same volume.

[14] Abner Shimony, 'Coherence and the Axioms of Confirmation', *Journal of Symbolic Logic* 20, #1 (1955), 1–28.

[15] Richard von Mises, *Probability, Statistics and Truth* (London: William Hodge and Co., 1939).

[16] Hans Reichenbach, *The Theory of Probability* (California: University of California Press, 1949).

[17] Ian Hacking, *Logic of Statistical Inference* (Cambridge: Cambridge University Press, 1965), p. 135.

[18] Pierre Simon de Laplace, *A Philosophical Essay on Probabilities*, trans. and rep. (New York: Dover Publications, 1952).

[19] The use of the calculus to extend the probability theory to include continuous

variables is discussed at the conclusion of this chapter.

[20] Thomas Bayes, 'An Essay Towards Solving a Problem in the Doctrine of Chances', *Philosophical Transactions of the Royal Society of London* 53 (1763), 370–418.

[21] There is no question but that Laplace knew that his Rule of Succession is a consequence of the Indifference Principle, under a tacit commitment to CC.

[22] George Boole, *The Laws of Thought*, rep. (New York: Dover Publications, 1951).

[23] John Venn, *The Logic of Chance*, rep. (New York: Chelsea Publishing Co., 1962).

[24] An exceptionally simple example is cited by J. Keynes, *Treatise of Probability*, 2nd ed. (New York: Macmillan and Co., 1921), p. 42; which he credits to von Kries.

[25] Harold Jeffreys, *Theory of Probability*, 3rd ed. (Oxford: Clarendon Press, 1967).

[26] The very recent work of D.A.S. Fraser, *The Structure of Inference* (New York: John Wiley and Sons, 1968), gives good chase. It is beyond the scope of this investigation to give his views on the fiducial argument an adequate hearing. However, Fraser's reliance on group theoretic symmetries does not avoid the problem discussed in Section 4.8. Also, see Appendix 9.3.

[27] Venn takes this result most casually, thus reaffirming his strict Frequentism. See: Venn, *The Logic of Chance*, Chapter VII, §15–16.

[28] The Total Evidence requirement (3) is *a methodological* principle, relevant to the application of inductive logic for a given agent at a specified time, and is assumed to apply for all theories considered. Also, (5) *objectivity* is a derived quality of the strict Frequentist position, since the probability$_2$ concept is point-valued when defined at all.

[29] In this work the Fisherian theory of likelihood ratio support, based on the likelihood function (defined at the end of this section), is discussed only briefly and then primarily with respect to Ian Hacking's reconstruction of the Fiducial argument (presented in the Chapter 6). The position taken here is that Fisherian likelihood can be seen as a central component in a theory of inductive *relevance*, and is discussed from this perspective in the final chapter, dealing with relevance and the programs of experimental design. This approach is intended not as a denial of the inductive support interpretation of likelihood, so often presented, e.g. G. Barnard's views, [for example, see, George Barnard, 'The Use of the Likelihood Function in Statistical Practice', *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* 1 (1965), 27–40] A.W.F. Edwards' work [A.W.F. Edwards, *Likelihood*, (Cambridge: Cambridge University Press, 1972] and, of course, Hacking's proposals; but as an alternative and mutually consistent exposition.

[30] Hacking, pp. 89–114.

[31] See Neyman, 'On the two different aspects of the representative method', *SESPJN*, paper #10, pp. 100–101, and Fisher's comments, pp. 136–138.

[32] We distinguish the functions (designated by X, Y, . . .) from their real values (designated by variables x, y, . . .).

[33] R. Carnap, *Logical Foundations of Probability*, 2nd ed., (Chicago: University of Chicago Press, 1967), p. 348.

[34] R.A. Fisher, 'On an Absolute Criterion for Fitting Frequency Curves', *Messenger of Mathematics* 41 (1912), 155–160.

[35] This definition is taken from D.V. Lindley, *Introduction to Probability and Statistics*, vol. 2 (Cambridge: Cambridge University Press, 1970), p. 46. Early work on sufficiency is due to Fisher, 'On the Mathematical Foundations of Theoretical Statistics', *Phil. Trans. Royal. Soc.*, A 222 (1922), 309–368. Also, in *CMS*, paper #10.

Fisher, 'Two New Properties of Mathematical Likelihood', *Proc. Roy. Soc., A* 144 (1934), 285–307. Also in *CMS*, paper #24.

The Neyman Factorization Theorem, giving both necessary and sufficient conditions for sufficiency dates from 1935: 'Su un teorema concernente le cosiddette statistiche sufficienti', *Giornale dell'Istituto Italiano degli Attuari*, VI, #4 (1935). Fisher's work gives only sufficient conditions. For additional commentary, see: J. Neyman and E.S. Pearson, 'Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses', *Statistical Research Memoirs* 1 (1936), 113–37. Also, rep. in *Joint Statistical Papers of J. Neyman and E.S. Pearson*, (California: University of California Press, 1967), paper #9. Henceforth, *JSP*.

[36] This definition is taken from Lindley, vol. 2, p. 57.

[37] Henceforth, variables with subscripts (usually) designate constants. Otherwise they stand for random variables on repeated trials of the same stochastic process. The context specifies which reading is appropriate.

[38] Alfréd Rényi, *Foundations of Probability* (San Francisco: Holden-Day, 1970).

[39] William Feller, *An Introduction to Probability Theory and its Applications* (New York: John Wiley and Sons, 1966), vol. 2, pp. 136–140.

[40] Lindley, vol. 2, p. 33.

CHAPTER 2

NEYMAN-PEARSON THEORY

In this chapter I examine the Neyman-Pearson theories of hypothesis testing and confidence intervals. Before considering the Fisherian program, I review the deficiencies that exist when the orthodox N-P program is interpreted as a *Frequentist* solution to inverse inference. Since my goal is to reconstruct Fisher's fiducial argument so that it supplies satisfactory inverse inferences, the problems with N-P theory serve as benchmarks for the proposed reconstruction of Chapters 3 to 5.

Section 2.1 focuses on the basic Frequentist difficulty in direct inference, namely, selection of the *reference class*. Neyman's pragmatic solution is compared to Reichenbach's epistemologically based recommendations and each is examined in the light of the *total evidence* requirement. This discussion permits reformulating I. Hacking's 'before-trial/after-trial' distinction in terms of the reference class problem. The section concludes with an outline of the argument for the remainder of the chapter.

Section 2.2 offers the reader a summary of the N-P theory of *hypothesis testing* with simple statistical alternatives. The concepts of *size* and *power* of a test in terms of $type_1$ and $type_2$ errors are rehearsed. The section ends with an illustration of the *Fundamental N-P Lemma*, which relates the *most powerful* tests to *likelihood ratio* tests.

Section 2.3 provides the reader with an examination of two surprising consequences of the theory developed in 2.2. By introducing a *randomizer* into the set-up, one can improve the power of the best test available if the data originally present arise from a discrete sample space of possible outcomes. That is, the randomizer acts to 'smooth out' the distribution. This is symptomatic of the fialure of N-P theory because the outcome of the randomizer is *ancillary* information for the hypothesis tested.

Second, using an example due to D.R. Cox, one sees how the *mixing* of two tests, each of optimum power, can result in an inferior test on net. This too results from the use of a randomizer to alter the set of possible outcomes in such a way as to allow a trade off between the two 'best' tests, exchanging size and power considerations between them.

Section 2.4 completes the review of N-P hypothesis testing with a discussion of the criteria used for evaluating tests with *composite* alternative hypotheses. The concepts of *bias* and *invariance* of a test are defined.

In *Logic of Statictical Inference*, I. Hacking argues that the N-P criteria for evaluating the merit of statistical tests are plausible on a 'before-trial' analysis, but may be worthless once particular outcomes become known. He bases this charge on an example, discussed in 2.5, showing that likelihood ratio tests can be biased, i.e., 'worse than useless.' I point out that Hacking's example shows only low level incompatibility between likelihood and N-P criteria, since Hacking uses *mixing*, *bias* and *invariance* considerations to identify the N-P preferred test. An alternative proposal is that Hacking's argument can be strengthened by using the deficiencies in N-P confidence interval theory instead of the hypothesis testing theory. This strategy is pursued in Section 2.7 and 2.8.

As a prelude to that investigation, in 2.6 I discuss the 'Stopping Rule' problem. Although the N-P Lemma (2.2) shows that the most powerful test of an hypothesis is a likelihood ratio test, the 'Stopping Rule' problem reveals that whether or not a particular outcome is part of the critical region (rejection region) depends upon the sample space of possible outcomes. Thus, in its sensitivity to alternative sampling plans (a 'forward looking' comparison), the N-P theory is in violation of a standard *Likelihood principle* (6.1), which is a consequence of *confirmational conditionalization* (condition (ii) of 1.2). Moreover, I note that this problem with the stopping rule for N-P theory suggests an objection which can be raised against the Fisherian arguments for randomization in experimental design, as Fisher uses a 'forward look' in defense of randomization. This idea is developed later on in the text (8.1).

Returning to themes presented in 2.5, Section 2.7 begins a review of Neyman's *Confidence Interval* theory. There I choose to emphasize the connections between the N-P criteria used to evaluate tests and the standards used to pick optimum confidence intervals. This prepares the way for an objection to N-P theory, voiced by way of an example, that good confidence intervals may have serious drawbacks when considered in light of particular data. Specifically, the text illustrates how 'shortest' confidence intervals (based on a *most powerful* test), may, at times, fail to exclude any permissible parameter value. Finally, we see that an interval system based on a maximally biased test (with zero power) may be used to correct this deficiency in the 'best' intervals.

The difficulties surrounding statistical inference involving more than one unknown factor are of continuing interest both for theoretical and practical reasons. Section 2.8 contains three maior components. [1] I propose a new constraint for *simultaneous confidence intervals*, to wit: *Confidence Equivalence*. Since functional relationships between parameters restrict the logically coherent answers one can give to questions raised simultaneously, e.g. $\theta_1 - \theta_2 = 0$ iff $\theta_1/\theta_2 = 1$, it is reasonable to constrain confidence intervals in order that these connections be respected. [2] In order to satisfy Confidence Equivalence, I suggest a technique for generating coherent sets of intervals (the *Master Test*), along with an illustration for the familiar bivariate normal model. [3] An important consequence of this exercise is the discovery that when satisfying the new condition on simultaneous intervals, one is forced to admit deficiencies of the very sort considered in Section 2.7 for the purpose of overturning the othodox theory.

The concluding section of the chapter illustrates the connection between hypothesis testing and interval theory by showing how the mixed test technique can be applied to confidence interval theory. Specifically, the infamous Behrens-Fisher problem is given an exact confidence interval solution with the aid of a randomizer. Historically, the Behrens-Fisher problem was the first to show the important differences between fiducial theory and Neyman's confidence intervals.

2.1. INTRODUCTION TO THE NEYMAN-PEARSON THEORY OF HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

Among twentieth-century theories of statistical inference, without doubt, the most widely used approach is due to the work of J. Neyman and E.S. Pearson.[1] The foundations for this program were set in the late 1920's and

early 1930's.[2] By the end of World War II the overwhelming majority of practising statisticians took it as the received position, and it has been only in the last two decades that serious rivals have emerged. Of these, two – the minimax program[3] (due to A. Wald) and the empirical Bayes methods[4] (due to H. Robbins) – are first generation descendants of the orthodox Neyman-Pearson theory and must be seen as purifications distilled from the original, primitive one.

The great popularity of the older N-P theory can be traced to three principle features of the program: The explicit frequentist basis – displayed in the interpretation of probability as probability$_2$. The clear mathematical formulation, especially of once vague testing and estimation problems. Third, the ease of application of the theory to many practical cases; most significantly those involving repeated testing or repeated sampling in a clearly specified sequence of trials.

For the discussion here, only a limited aspect of this program will be reviewed. The two major conclusions argued for are: (1) that the theory of hypothesis testing fits neatly into the theory of confidence intervals, and (2) neither of these are adequate for a theory of inference, especially a theory of *inverse* inference. Both these results are well known, and in fact were often voiced by Neyman himself. The focus for the analysis is the difference between those statements (of probability) made before trials are run, i.e. before knowing the sample data, compared to those made afterwards. In this respect the fundamentals of the discussion follow closely that given by I. Hacking in his *Logic of Statistical Inference*.[5]

The specific points of the argument include: the consequences of using a strict probability$_2$ interpretation as it affects the determination of the 'reference class' for direct inference; an evaluation of the insensitivity of the operating characteristic criteria, e.g. power, to modifications in the information available and finally, a minor suggestion how some of the apparently less desirable results of the previous analysis might be converted into an advantage through an additional consistency principle, called Confidence Equivalence.

A good starting place for the discussion of the Neyman-Pearson theory is with a brief digression concerning a typical Frequentist problem: the determination of the 'reference class' for the direct inference. The simplest formulation of the difficulty is that a single chance event can be the outcome of more than one stochastic process. Hence, without further guidance, the Frequency Principle can suggest alternative, incompatible probability$_1$ values for the same event; one using one stochastic process with its probability$_2$

and the second based on a different probability$_2$. For example, the probability that the next flip of this 'fair' coin landing 'heads' is $\frac{1}{2}$, by the Frequency Principle. However, it might also be the case that the coin was obtained from an urn containing half fair coins and half biassed for heads, say $\theta = .7$, in which case the probability$_2$ for a 'heads-up' flip on a composite trial of selecting a coin at random from the urn and flipping it is .6, and not .5. The next flip is an outcome of both kinds of stochastic processes. The question is, which (if either) of the two probability$_2$'s is the legitimate evidence for using the Frequency Principle?

A solution, credited to Reichenbach, requires that the evidence used for the Frequency Principle refer to the *narrowest* reference class for which precise information is available.[6] This is explicated as follows: For Reichenbach, as for many other Frequentists, probability (interpreted as probability$_2$) does not apply to single events directly. Instead, the chance event is located in a sequence of other chance events, supposedly related by the 'stochastic' property common to all. This sequence constitutes the 'reference class' for the event in question. The Reichenbachian requirement is that the Frequency Principle (which for him determines a 'weight' in place of a probability$_1$ statement) be based on a sequence or reference class which is a sub-sequence or sub-class of all those made available through probability$_2$ evidence known to the investigator.

For instance, referring to the coin case (above), the probability$_2$ fact that the coin is 'fair' translates (for Reichenbach) into a sequence of hypothetical or 'possible' flips of this coin (of which the *next* flip is a member). It is supposed that the sequence has the mathematical property that, in the limit, the relative frequency of heads equals $\frac{1}{2}$; thus, the 'limit of relative frequency' interpretation for the probability$_2$ concept. Similarly, the composite trial of selecting a coin from the urn and then flipping it generates a sequence in which the limit of frequency for 'heads' is .6. However, since the next flip (with this coin) is an instance of both kinds of trials, the sequence of flips with this very coin preceded by drawing this coin from the urn forms a sequence of boht initial sequences. Moreover, this new sequence has the limiting frequence value .5 for 'heads' (as it is a proper sub-sequence of the sequence of flips with this coin). Therefore, using the Reichenbach rule for selecting the narrowest 'reference class' to determine the probability$_2$ for applying the Frequency Principle, yields the value $\frac{1}{2}$ for the probability$_1$ (or 'weight') that the next flip of this fair coin (selected from the urn) will land 'heads-up.'

Although Reichenbach's regulation is intended to apply to an interpretation

of probability$_2$ as a limit of relative frequency in an infinite sequence, certain key features of his solution are generalizable over a larger domain of frequency interpretations. Most centrally, the 'narrowest reference class' rule is obviously an *epistemological* constraint, as it depends on what information is available to the investigator at the moment of its application. In this guise, it is a specialization of the Total Relevant Evidence regulation; demanding that all information about the event in question be used to determine the probability$_1$ assignment. For instance, since the value recommended by Reichenbach's rule for the coin case is .5, which equals the value based on the sequence of repeated flips of the 'fair' coin, the extra information about the stochastic genesis of the coin, i.e. that it came from the urn, is *irrelevant* to the probability$_1$, given the knowledge that the coin is 'fair'. Reichenbach follows a strict Frequentist program, so that the derived irrelevance of the information about the origin of the coin is based on a *stochastic* or probability$_2$ irrelevance of the same sort. That is, it is probabilistically$_2$ irrelevant to a flip of this fair coin that it was selected from the urn before flipping. This last claim is supposed to convey an empirical fact which serves as the basis of the consequent probability$_1$ irrelevance. The empirical fact is, of course, that the limit of relevance frequency for 'heads' is the same in the original sequence of repeated flips with this coin and in the sub-sequence of flips preceded by drawing it from the urn.[7]

Among empirical, frequentist approaches, Neyman's is almost disarmingly simple although the simplicity can also be misleading. A good part of the difficulty stems from Neyman's own attempt at explication by way of the 'classical' or Laplacean 'equal probability' interpretation.[8] He argues that the probability$_2$ of an object A possessing the property B is the ratio of the number of $(A \,\&\, B)$ objects to the number of A objects. In this definition, the set of A objects is called the *Fundamental Probability Set* [FPS] and is *intended* to duplicate the reference class of the probability statement. If the FPS is finite, then the ratio is just the ratio of the two cardinal numbers for the classes involved. If the FPS is infinite, specifically, a region of n-dimensional space, then a (Lesbesgue) measure over the reference class replaces the cardinality measure. It is supposed then that subsets of the FPS, e.g. B (above), are all measurable.

Unfortunately, this interpretation suffers from two glaring problems. How is the FPS determined? How is the measure assigned for the continuous FPS? More confusing is Neyman's common use of the sample space or set of 'possible' outcomes for the FPS. This would, following the rule above, lead to an equal or uniform probability distribution automatically. For example,

knowing that a coin can, when flipped, land either 'heads' or 'tails' uppermost would, by the Neyman rule, lead to a probability$_2$ of .5 for a flip having either outcome. But this is acceptable only for a 'fair' coin.

The solution to this problem, that is the determination of the FPS along with associated measures, gives Neyman's answer to the direct inference problem: How to fix the reference class? Unlike Reichenbach's response, involving a clearly epistemological criterion, Neyman's suggestion is *pragmatic*. The *pragmatic* component is the determination of a sequence of 'repeated trials' along with a hypothesized frequency. Instead of requiring that the sequence reflect the investigator's knowledge about the specific trial in question, e.g. the *next* flip of the coin, it is determined by a practical concern. Paradigmatic cases involve a *routine* for generating the repeated trials, e.g. quality control of batches of goods produced by a certain machine.

Thus, the FPS may just as well be the sample space of 'possible' outcomes so long as the associated measure is derived from the sequence of 'repeated trials.' For those cases where the routine is well defined, the measure reflects the *empirical* content of the probability$_2$ term. However, even knowing the routine does not guarantee that *knowledge* of the empirical frequencies is also available. Neyman removes this last difficulty by appeal to the so called 'law of big numbers.' It is supposed that, given the routine for generating the repeated trials, e.g. repeated flips with this very same coin, and statistical knowledge of the right sort, e.g. that the coin is fair, gives a high probability$_2$ (interpreted now as 'practical certainty') for a narrow range of empirical frequencies in the sequence of repeated trials. That is, it is claimed to be practically certain that in a long sequence of repeated flips of this coin the observed frequencies approximate the 50%–50% proportion for 'heads' and 'tails' outcomes. This gives the 'repeated sampling' property of the direct inference under the Neyman interpretation. When such a 'repeated sampling' holds, the single trial is said to be a 'random' sample from the FPS with associated measure.

This idea of randomness as the connecting link between the measure of the FPS and the 'repeated sampling' property, necessary for any direct inference solution, supplies the missing ingredient for cases where the routine is *not* clearly given by an independent concern. For these cases, loosely labelled 'cognitive,' the investigator can work backwards from a claim of randomness to a hypothetical reference class generated by a postulated routine. To the degree that the imagined 'repeated trials,' so constructed, approximate a routine that might be actualized, the derived frequencies have *empirical* content. Therefore, there is much room for disagreement in the more 'cognitive' problem.

The following remarks (of Neyman) support this reconstruction.

> The justification of the way of speaking about the definition of the measure within the fundamental probability set in terms of imaginary random experiments lies in the empirical fact, which Bortkiewicz insisted on calling the law of big numbers. This is that, given a purely mathematical definition of a probability set including the appropriate measure, we are able to construct a real experiment, possible to carry out in any laboratory, with a certain range of possible results and such that if it is repeated many times, the relative frequencies of these results and their different combinations in small series approach closely the values of probabilities as calculated from the definition of the fundamental probability set. . . .
>
> These examples show that the random experiments corresponding in the sense described to mathematical defined probability sets are possible. However, frequently they are technically different, e.g., if we take any coin and toss it many times, it is very probable that the frequency of heads will not approach $\frac{1}{2}$. To get this result, we must select what could be called a well-balanced coin and we have to work out an appropriate method of tossing. Whenever we succeed in arranging the technique of a random experiment, say E, such that the relative frequencies of its different results in long series sufficiently approach, in our opinion, the probabilities calculated from a fundamental probability set (A), we shall say that the set (A) adequately represents the method of carrying out the experiment E. The theory developed below is entirely independent of whether the law of big numbers holds good or not. But the applications of the theory do depend on the assumption that it is valid.[9, 10]

In summary then, the application of the mathematical probability to an experimental situation requires (at a minimum) that there be a high probability$_{(1\ \&\ 2)}$ that the observed frequencies on repeated trials of a (hypothetical) experiment be approximately those given by the mathematical probability. Where the problem is 'practical' the repeated sampling occurs within the sequence fixed by the routine, e.g. the sequence of batches of goods produced by this machine. Where the problem is 'cognitive' the repeated sampling occurs within the sequence fixed by the empirical claim that the sample is 'random,' given the mathematical probability law. For both cases, once the sequence or routine is determined (thereby fixing the measure for the FPS) the probability for a single event or object is also determined – *no matter what else is discovered about the instance at hand.* Thus, the direct inference problem is handled without requiring that all the investigator's background knowledge be cited; not even all of it which, intuitively, might be relevant to the single trial.

By way of illustration of this consequence, it is important to review the distinction (for Neyman) between random variables and unknown constants. This difference reflects, in another way, the separation between direct inference and inverse inference. Neyman defines a *random variable* as,

a single-valued measurable function (not a constant) defined within the fundamental probability set (A), with the exception perhaps of a set of elements of measure zero.[11]

Therefore, for arbitrary subsets B of the FPS (the sample space), a random variable x can be defined so that instead of using the phrase 'an A being a B,' one can substitute 'x falling the range $[\underline{B}, \bar{B}]$'; where 'x' is a variable for the experimental outcome in question.

Suppose that the 1000th digit in the decimal expansion of π is unknown. Clearly, it is either equal to 7 or not. Moreover, there is no stochastic ingredient in the determination of the mathematical fact. That is, there is no 'chance' element involved in the truth (or falsity) of the mathematical claim that it equals 7. Hence, according to Neyman, any FPS consistent with the mathematics of the problem must assign the (unknown) constant value to the 1000th digit in the decimal expansion of π. By the definition (above) the 1000th digit is *not* a random variable.[12]

Consider, next, a modified problem, described as follows: Construct a FPS of the ten values $(0, \ldots, 9)$ and assign a measure m in accord with the rule: The value $\mathbf{m}(i)$ $(i = 0, \ldots, 9)$ equals the proportion of digits equal to i in the first million of the decimal expansion of π. Finally, let R be a method for selecting digits from the first million in the decimal expansion of π so that the foregoing FPS (and measure) applies, i.e. so that the repeated sampling property holds for R. One can then say that the probability of a digit selected by R being a 7 is $\mathbf{m}(7)$; where $\mathbf{m}(\ )$ is the measure over the FPS. Specifically, this probability applies to the *next* digit selected by rule R. Imagine, however, that it becomes known that, in fact, the next digit produced by R is the 1000th. Since the probability is determined by the FPS (and measure), it remains $\mathbf{m}(7)$ for the claim that the next digit selected is a 7 – even knowing that it is the 1000th. This is because the routine by which the measure m is defined makes use of repeated application of R, and does not take into account subsequences of sub-routines of R *known* to contain the single trial (the next one) up for consideration.

Neyman's remarks about these sorts of cases often revert to the claim that the probabilities for unknown constants (as opposed to random varaibles) are *intermediate*, i.e. either 0 or 1, but unknown.[13] This is somewhat different from the conclusion suggested in the example above, where the probability that the 1000th digit of π (a constant) equals 7 is *undefined* because there is no stochastic factor to appeal to. The two attitudes can be reconciled when the following maneuver is available. Under the second case (above) there is a stochastic ingredient or probability$_2$ for the selection of the 1000th digit of π using R (by assumption, it equals $1/10^6$). Also, there is a

probability$_2$ that the digit produced by R is a 7. Therefore, by the probability calculus, there is a probability$_2$ that, conditioning on the digit being the 1000th, it is also a 7. This value is 1, if true, and 0, if false, that the 1000th digit is a 7.

This concludes the preliminary discussion of the Neyman-Pearson solution to the direct inference problem. The sections following examine the principal features of their well known views on hypothesis testing and confidence intervals. These supply the strict frequentist response to the demand for some sort of 'inverse' conclusions. Since prior probabilities are unavailable (for a variety of reasons discussed below), posterior conclusions cannot be reached. Instead, the N-P theory makes use of 'tests' with critical controls expressed by way of two sorts of 'errors' than can result from the application of the 'tests.'

The stochastic or probabilistic element is incorporated into the analysis by calculating the probabilities for the various types of errors associated with a particular test. These error probabilities are, of course, based on a repeated sampling interpretation of some FPS with affiliated measure. As in the example above, it is imperative that careful attention be paid to the *routine* so as not to (mistakenly) attempt to include additional information in the determination of the error probabilities. More precisely, the probabilities must *not* become altered because of knowledge available about some specific trial, e.g. the next one, which is *not* true of all trials in the repeated trial sequence.

This proviso is, henceforth, to be known as the 'before trial – after trial' or 'forward look – backward look' condition. Following I. Hacking's examination of the N-P theory,[14] it will be argued that the program has plausibility *only* on the 'forward look' or 'before trial' perspective. However (unlike Hacking's) this argument is pressed against the confidence interval portion of the N-P theory, and therefore, only indirectly against the hypothesis testing half of their program. Part of the difficulty with Hacking's strategy, which attacks hypothesis testing directly, is that for the most elementary and intuitively clear cases the N-P solution (a *most powerful* test) has favorable 'after trial' or 'backward looking' consequences. In fact, they are in perfect agreement with the tests offered by Hacking as replacements (*likelihood ratio* tests).[15] It is only when the N-P program resorts to less than first rate error considerations, e.g. *invariance*, that Hacking is able to clearly demonstrate a clash between forward and backward looking analyses, and then only through an example which relies, *essentially*, on a small sample size.[16] This remark is not intended as a denial of the validity of Hacking's criticism. However, it seems desirable to reinforce or bolster the

argument by showing that it is not restricted to tests based on less than 'best' error evaluators, or just small samples.

To recast the dialectic, a committed Neyman-Pearsonite might accept Hacking's examples as evidence for a need to review the lower level N-P error criteria, especially with small samples. If Hacking's own suggestions are not acceptable as a resolution to the problem, then the statistician might argue for a *suspension of judgement* in the controversial cases; in the expectation that with increasing data the conflict might disappear.

In the hope of eliminating even this escape route from the 'after trial' pitfall, the following line of attack is adopted: First, the traditional hypothesis testing theory is reviewed, including several more controversial tests involving randomization. Then, using the hypothesis testing theory as a foundation, Neyman's theory of confidence intervals is reconstructed. The important point is that confidence intervals are derivative from hypothesis tests. The best confidence intervals are generated from the best hypothesis tests. Third, it is shown that the *best* confidence intervals suffer from the after-trial deficiency that, sometimes, i.e. for some data, they are *logically trivial*, i.e. tautologous. For emphasis, it is also shown that *inferior* hypothesis tests, can have *better* after-trial consequences. Therefore, it is claimed that the N-P theory cannot serve as an adequate replacement for an inductive logic, especially for inverse inference.

Finally, in mitigation of this last unfavorable result, a proposal is made for improving the N-P theory. The suggestion is that a *new* condition, called Confidence Equivalence, be adopted in addition to the other error evaluators. Confidence Equivalence [CE] knits together different tests, and therefore different confidence intervals, while securing logical consistency (of a sort). The application of CE is demonstrated through a Master Significance Test [MST] for the bivariate normal distribution. It is shown that the resulting intervals, while obeying CE, can suffer from the logical triviality described above. Specifically, the solution of the ratio of normal means problem (due to Fieller) has the dual property of obeying CE while being logically trivial for some data. Of course, this does not fully excuse Fieller's solution from the tautologous intervals it generates. But the CE property does supply some novel defensive arguments in favor of these types of otherwise unsatisfactory (on the 'after-trial' analysis) tests and intervals.

2.2. HYPOTHESIS TESTING

Perhaps the most elementary version of inverse inference concerns the use of a sample to reflect some trait of the more general population from which

the sample comes. For example, suppose that a well defined finite population is examined to determine the proportion having some property F. It is practically impossible to exhaustively survey the population at large, so that a sample is taken according to some rule Q. Assume, also, that the magnitude of the sample is very much less than that of the population.

A simple application of the Neyman-Pearson theory of probability for this case is to hypothesize a probability law for the experiment according to which the FPS consists of n-fold selections from the parent population where n is the sample size) and the probability of a member of the sample having F is just the unknown proportion of F's in the parent population. It is also claimed that sampling according to Q generates 'random' samples, which can be considered as sampling individuals 'with replacement,' given the relative magnitudes of the sample and parent populations.

In other words, a *mathematical model*, stochastic in nature, can be supplied for the survey problem such that the probability of a sampled entity being F is θ, where θ is the unknown proportion of F's in the parent population. If, as supposed, the magnitude of the sample is much less than that of the population at large, then a model using sampling replacement (after each entity is selected) is a good approximation for rule Q. Thus, through the mathematical model, the hypothesis that $\theta = k$ (k a given constant) is converted into an elementary statistical hypothesis; to wit, that the probability for F's in the sampled group is binomial, with parametric value for F equal to k. Of course, the adequacy of the mathematical model (apart from the speculation that $\theta = k$) rests heavily on the translation form, 'The sample is selected by rule Q.' to, 'The sample is random.' To repeat, this translation is verified in any practical case by appealing to the 'repeated sampling' frequency properties of rule Q in a hypothetical experiment.

If, in fact, the model is accepted and the hypothesis true, then probabilities for random variables calculated from the sample, e.g. the sample proportion of F's (denoted $\bar{x}$), are also available. It is the basic conjecture of the Neyman-Pearson program that these random variables can serve as relevant evidence for testing the hypothesis $h_0 : \theta = k$; given the binomial model using rule Q. The pragmatic context of *hypothesis testing* is given by Neyman as follows:

> The problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B, with no other course of action contemplated. Moreover, in order to speak of a test of a statistical hypothesis, it is necessary that the desirability of actions A and B depend on the frequency function $p(e)$ of some observable random variables and that $p(e)$ be uncertain.[17]

The first point to be established is that, by itself, the calculus of probabilities is of little value for resolving the uncertainty. The mathematical model by which the hypothesis becomes statistical fixes the FPS so that θ, although unknown, remains an unknown constant and not a random variable. Therefore, one cannot sensibly speak of the probability that, say $k_1 \leqslant \theta \leqslant k_2$, since (given the FPS from the model) it is undefined or indeterminate: as it equals either 1 (if the inequality is true) or 0 (if it is false).

In and of itself, the existence of the simple binomial model does not preclude the application of yet another, more complex one in which θ can be construed as a random variable. However, in the cases considered, no such alternative model is forthcoming because either (i) it is known that θ does not have a stochastic origin, e.g. it is known that the parent population was determined in a non-stochastic manner, or (ii) even if the genesis can be presumed to be stochastic, e.g. the population was selected according to some chance mechanism, nothing is known about the underlying chance device so that no useful or stochastically informative model can be constructed. Both cases are lumped together under the heading, 'No frequency based prior probabilities available for the unknown parameter.'

Thus, the calculus does not permit the determination of the probability that, for instance, action A will be favourable when $\bar{x} = \bar{x}_0$ (the observed value of the random variable). Specifically, inverse inference is blocked: where 'action A' is interpreted as 'believing h_0' and 'desirability' is interpreted as 'is true.' In other words, one cannot determine the probability that h_0 is true, given $\bar{x} = \bar{x}_0$ and the FPS derived from the mathematical model.

In place of a test interpretable as an inference (for or against h_0), the theory considers a *rule of inductive behaviour*, encapsulated in a *test of a statistical hypothesis*.

DEFINITION: In the above circumstances, any rule R prescribing that we take action A when the sample point E determined by observation falls within a specified category of points, and that we take B in all other cases, is a *test of a statistical hypothesis*.[18] That is, any rule which determines the course A or B depending on the observation of a random variable whose probability is a function of the hypothesis under investigation, is a test of that hypothesis, i.e. constitutes a rule of inductive behaviour.

In general, it is assumed that the desirability of each option (A or B) is known once the truth of the hypothesis is determined. If A is the more favourable one when h_0 is true (and, by assumption, the less favourable when h_0 is false), then all sample points which lead to action B are labeled 'rejection

values for h_0,' and the set of such points are denoted the 'rejection region' or 'critical region' for h_0, subject to the test rule considered. Furthermore, if the rule is adopted, then h_0 is said to be 'rejected' if the random variable falls in the critical region.

The Neyman-Pearson program for evaluating alternative rules of inductive behaviour is based on the comparison of two sorts of 'error' associated with any test. An 'error of type$_1$' occurs when course B is adopted and h_0 is true. In other words, a type$_1$ error results from 'falsely rejecting h_0.' An 'error of type$_2$' occurs when course A is adopted and h_0 is false. In other words, a type$_2$ error results from 'falsely accepting h_0,' i.e. failing to reject h_0 when it is not true.

The comparison is effected by calculating the probabilities for each of the two errors, assuming the mathematical model. The maximum probability for a type$_1$ error is labeled the 'size' of the test, and is denoted by the letter 'α'. Type$_2$ errors are captured by the 'power' of the test, which equals 1 minus the maximum probability of a type$_2$ error, and is denoted by the letter 'β'. Intuitively, good tests are those of low size and high power. The connection between size, power and empirical frequencies, rests on the ability to actualize the hypothetical experiment used to justify the application of the mathematical model to the problem at hand. That is, using the law of big numbers, it is 'practically certain' that the frequencies of the two errors resulting from applying the statistical test to a sequence of 'random' experiments will agree (in the long run) with the values α and β. In the long run, the proportion of 'false rejections' of h_0 will be α (at most) and the proportion of 'false acceptances' of h_0 will be $1-\beta$ (at most).[19] These frequencies are labeled the 'operating characteristics' of the test.

Although the size, α, of a test is specified once the hypothesis h_0 is stated, the power depends on just what alternatives to h_0 are possible. The alternatives are determined by two general features. First, the adoption of a stochastic mathematical model circumscribes, in broad terms, a range of alternatives. For example, the use of a binomial model for the population survey problem (under rule Q) delimits a one-parameter context, whose values are confined to the [0, 1] range (representing the possible proportions for F-ness). Loosely speaking, the model captures the number of relevant factors involved in the problem through the number of parameters introduced. Second, more specific knowledge of the problem situation, more idiosyncratic than a 'model,' is reflected in a specification of the range of the parameter short of its potential values compatible with the model. For example, it might be known that, for the population sampled, the proportion of F-ness is bounded

by the values $\frac{1}{4}$ and $\frac{3}{4}$. This information is translated into the mathematical formulation of the problem by considering alternatives to h_0 which specify that $\frac{1}{4} \leqslant \theta \leqslant \frac{3}{4}$ $(\theta \neq k)$. This two level distinction is not sharp, since a two-parameter problem can be converted into a one-parameter case by restricting one of the factors to a constant value. However, in general, a change in parameterization represents a change in model, and not just a restriction on one of the parameters.

A statistical hypothesis is termed *simple* if it generates an exact probability for the FPS, i.e. the measures over the FPS are precise.[20] A statistical hypothesis which is not simple is called *composite*, and represents a 'disjunction' of many simple hypotheses. For one-parameter problems, simple hypotheses postulate a unique value for the unknown parameter, e.g. $\theta = k$. Similarly, composite hypotheses may posit only a range $[k_1, k_2]$ for the value of the parameter. By way of introducing the general Neyman-Pearson theory of hypothesis testing, consider first an elementary case where both the hypothesis being tested, and its sole alternative are simple. For illustration, suppose the population survey problem as before, with $h_0: \theta = \frac{1}{2}$ and simple alternative hypothesis, $\theta = \frac{1}{4}$.

Two fundamental restrictions on comparisons of statistical tests are presupposed. It is required that tests be compared only if they are tests of the same statistical hypothesis and also of the same size, i.e. for a common h_0 and with equal α value. This two-fold condition is based on the belief that errors of the first kind differ in importance from and cannot (in general) be traded-off against type$_2$ errors. The methodological rule of thumb is that errors of the first kind are more serious – are more important to avoid: so that h_0 is selected with this condition in mind.

The α level is fixed (presumably) by the pragmatics of the decision context. That is, depending on the consequences of adopting action B when h_0 is true (type$_1$ error), the investigator sets an α value compatible with the needs at hand. Conventional α's (size) are .1, .05, .01 and .001. The relatively small magnitudes reflect the seriousness of the type$_1$ errors, and *not* a statistical or probabilistic restriction on the construciton of tests.

With α and h_0 fixed, different tests are ranked by their corresponding power. The greater the power the less the probability of committing an error of the second kind; hence the greater the sensitivity or resolving ability of the test. That is, tests with greater power are more likely to reject h_0 when it is, in fact, false. The most important result of the Neyman-Pearson theory is that for tests of a simple hypothesis against a simple alternative (assuming continuous random variables), there exists a best or most powerful test (for

any α). More importantly, the best test is also a likelihood ratio test; which is defined as follows.

Let x be the random variable for which the test is defined. Then, let $\lambda = p(x/h_0)/p(x/\sim h_0)$, the ratio of the probabilities for the random variable under the null hypothesis h_0 and its simple alternative $\sim h_0$. Let the critical region for h_0 (of size α) C_λ be defined as the set of x values, of total probability α, with minimum λ values. That is, h_0 is rejected so long as the likelihood ratio is low. Then, the best test for h_0, against the simple alternative, is just the likelihood ratio test.[21]

To continue with the example, if the sample size is n (n individuals selected according to rule Q) then the probability that r of them will show F-ness is:

$$C_n^r \cdot \theta^r \cdot (1-\theta)^{n-r};$$

where C_n^r is the binomial coefficient equal to $n!/r! \cdot (n-r)!$, and θ is the unknown proportion of F's. Therefore, the likelihood ratio for h_0 against the simple alternative equals:

$$\lambda = (\tfrac{1}{2})^n/[(\tfrac{1}{4})^r \cdot (\tfrac{3}{4})^{n-r}].$$

λ is a strictly increasing function of r, so that the C_λ test for h_0 (against the simple alternative $\theta = \frac{1}{4}$) uses low values of r for rejection. If the random variable $\bar{x} = r/n$ (the sample proportion of F's) is used, C_λ utilizes a 'one-tailed' rejection region for h_0 defined as:

$$C_\lambda = \{\bar{x}: \bar{x} \leqslant \kappa\};$$

where κ is a constant determined by α, the size of the test, alone. By the Neyman-Pearson lemma, it is also the most powerful test for h_0.[22]

2.3. TWO USES OF RANDOMIZERS TO PRODUCE 'MIXED' TESTS WITH IMPROVED POWER OVER 'PURE' TESTS

The following two examples show up some of the less intuitive ways the likelihood ratio rule for most powerful tests can be utilized, in conjunction with the 'forward look,' to achieve superior tests. The key is the addition of a 'randomizer' with known stochastic properties independent of the unknown parameter. In the first case the randomizer is described explicitly. In the second, it is built into the problem by supposing that one feature of the experiment is the random selection of a measuring apparatus. Using the randomizer changes the FPS from the more intuitive one, which is available without appealing to the randomization.

[I] Although the likelihood ratio test turns out to be best in terms of power, its applicability can be severely restricted for discrete distributions, due to the fixed size. For example, suppose that the random variable x can assume only 11 values, $(x_0, x_1, \ldots, x_{10})$, with probability $p(x_n)$, given either h_0 or $\sim h_0$, as displayed in the table below.

	$p(x_0)$	$p(x_n)$ $[n = 1, 2, \ldots, 10]$
h_0	.05	$n \cdot .019/1.1$
$\sim h_0$	.5	$n \cdot .01/1.1$

If α is set at a little less than .02, then the only possible test for h_0 that does not exceed the size limitation uses the critical region consisting of the value $[x_1]$. However, the likelihood ratio for x_1 (in fact for all x_n $[n = 1, 2, \ldots, 10]$) is .019/.01, while that for x_0 is .05/.5, or fully 1/19th that of the other λ values. If it were not for the exceptionally small test size, $\{x_0\}$ would be used as the critical region. That is, because of the α value, the only admissible likelihood test fails to make use of the region of the sample space with the lowest likelihood ratio.

This defect can be overcome by utilizing a 'randomizer,' in effect, to 'smooth out' the discrete distribution above into a relatively continuous one. For example, if the x reading is augmented by a y reading of a random variable uniformly distributed on [0, 1], then by enlarging the sample space to (x, y) a more powerful test is available. That is, by using the conjunction of x_0 with a set of y readings, say $y \in [0, .4]$ to form the critical region, a test with size $\alpha = .02$ and increased power is created. To repeat, the improvement represents the use of a critical region with better likelihood ratio, based on a randomizer to duplicate a continuous distribution. Of course, just which region of [0, 1] is used to form the rejection component of y does not matter, so long as it has a total probability of .4. Thus, knowing that $x = x_0$, on a particular trial, we see that the deciding factor for or against rejection of h_0 is merely the outcome of the randomizer. Since the probabilities are defined on a FPS which now includes the two-dimensional random variable (x, y), one can 'explain away' the counterintuitive result that the randomizer yields data relevant to deciding whether or not the reject h_0 by referring to the FPS used for the test. In other words, the test using the randomizer appeals to a different FPS, so that there is no conflict between the pure test and the 'mixed' or randomized one.

[II] A second use of randomization for improving statistical tests, based on

the Neyman-Pearson criteria of size and power, stems from the ability to combine two different tests into a single one, which has better operating characteristics than the simple combination of the two 'best' tests, taken singly. That is, if two tests are available, based on different experimental data, then by using a randomizer to combine the separate tests into a single one, say by picking an experiment at random, then the best overall test may be different from the simple conjunction of the best tests, when the experiments are considered separately.

An elegant example of this phenomenon is due to D.R. Cox, and is reported in H.E. Kyburg's *The Logical Foundations of Statistical Inference.*[23] Suppose that two experiments are available for testing the hypothesis $h_0: \mu = 0$ against the simple alternative $h': \mu = 10$; where μ is the unknown mean of a normal distribution whose variance, σ^2 is known, but which depends on the instrument used for measurement. For experimental set-up One $\sigma_1^2 = 1$, while for the other it is much larger, $\sigma_2^2 = 100$.

Given each experiment there is a best (most powerful) test for h_0 (at any α level), assuming the data to be a reading of the random variable x, which is distributed according to the normal distribution $N(\mu, \sigma_i^2)$ $(i = 1, 2)$. For instance, if $\alpha = .05$, then the most powerful test (given a single experiment) uses the critical region defined by:

$$C_\lambda = \{x: x > 1.64 \cdot \sigma_i\}.$$

However, if the two tests are combined into a single one, say by using a randomizer to pick i at random, i.e. $\mathbf{p}(i = 1) = \mathbf{p}(i = 2) = \frac{1}{2}$, then the best overall test for any α, is *not* based on C_λ (above). Note that the two tests are not of the same power. Experiment one is much more precise, hence with greater power. Therefore, if the mixed test is defined as:

$$C_M = \left\{x: \begin{array}{ll} x > 1.28 \cdot \sigma_i & \text{if } i = 2 \\ x > 5 \cdot \sigma_i & \text{if } i = 1 \end{array}\right\}$$

the difference in power of the two initial tests can be converted, to advantage, into an unbalanced single test. When $i = 1$, the chance of an error type$_1$ is much less than .05, very nearly zero in fact. However, when $i = 2$ there is just about a .10 chance of committing a type$_1$ error. Because the randomizer selects the second set-up only 50% of the time, the overall size is .05. Similarly, the power is not shared equally when $i = 1$ and $i = 2$. The increased 'size' for the $i = 2$ case carried an increase in power with it; whereas the decrease in 'size' for the case $i = 1$ hardly affects the power, since experiment one is already so precise. The net effect is a mixed test with better operating

characteristics than the intuitive C_λ test. Once again, the apparent conflict is dispelled by the reminder that, for the mixed test, a different FPS is employed. If the single experimental set-up One is used, then its FPS makes no reference to a random variable with $\sigma^2 = 100$, which arises from the second experiment. Similarly, with the second experiment the FPS makes no reference to a random variable with $\sigma^2 = 1$. However, in the mixed test the FPS includes both sorts of readings.

2.4. TESTS WITH COMPOSITE ALTERNATIVE HYPOTHESES

Having reviewed the principal features of the N-P theory for testing a simple hypothesis against a simple alternative, we see that the next stage is to consider tests of simple hypotheses against composite alternatives. For example, to continue with the population sampling problem, suppose that it is known that $\theta \leqslant .5$. Then, alternatives to $h_0 : \theta = .5$, consist of a continuum of simple hypotheses whose 'disjunction' is captured in the composite hypothesis: $\sim h_0 : \theta < .5$.

For any proposed test, and critical region C, the size remains exact – since h_0 is simple. However, the probability of an error type$_2$ may well vary with the true value of the unknown parameter(s). Instead of reporting just the maximum probability for error type$_2$, given $\sim h_0$, define the β-function (for the region C) as:

$$\beta(C, \theta) = \text{probability for the region } C, \text{ given } \theta.$$

The β-function records the power of the test for each simple alternative. The power for the test, defined as one minus the maximum probability of an error type$_2$, is then just $\min_{\theta \in \sim h_0} \beta(C, \theta)$. Under very special conditions, it sometimes happens that a test can be found such that its β-function represents the very best that can be obtained (at the assigned α level) for each simple alternative element in $\sim h_0$. In other words, C represents both a test of maximum power for h_0, and also the most powerful test for h_0 against any of the simple alternatives in $\sim h_0$. Such tests are called uniformly most powerful [UMP] tests.[24]

The connection between UMP-tests and C_λ, or likelihood tests, is quite simple. For composite alternative hypotheses, extend the definition of the likelihood test as follows. Assume that $\sim h_0$ is equivalent to the statement that θ, the unknown parameter(s), falls in the set of values denoted by Δ. Then,

$$\lambda_\Delta = \mathbf{p}(x/h_0)/\max_{\theta \in \Delta} \mathbf{p}(x/\theta).$$

The C_{λ_Δ}-test is defined, analogously, as the set of values of the random variable x, with total probability (given h_0 and assuming continuity of x) equal to α, such that the λ_Δ values are minimized. Then, if a UMP-test exists, it is a C_{λ_Δ}-test. By the N-P lemma for simple alternatives, it is also a C_λ-test for each of the simple alternatives in $\sim h_0$.

By way of illustration, there exists a UMP-test for the simple null hypothesis h_0: $\theta = \frac{1}{2}$ against the composite alternative: $\theta < \frac{1}{2}$, of the population survey problem. For each simple alternative the most powerful test uses a 'one-tailed' critical region, based on the sample proportion of F's, $\bar{x}$. In each case it is a 'left-hand tail.' This is reflected in the β-function, which is strictly increasing for decreasing values of θ.[25]

The existence of UMP-tests are severely limited, not merely by the connection with sufficient statistics (see note above), but more generally in the conditions set forth by Neyman.[26] By way of illustration of these meager conditions, if the population survey problem is altered so that the set of alternatives includes values for θ both greater than and less than the value posited by h_0, then no UMP-test exists (for any α). For one set of alternatives (the less-than group) the UMP-test is a 'left-hand tail' test, using $\bar{x}$; while for the other set it is a 'right-hand tail' test, again using $\bar{x}$.

Where no UMP-tests are available, the N-P theory resorts to secondary, and tertiary criteria: 'unbiased tests' and 'invariant tests.'[27] A test is *unbiased* so long as its power exceeds its size. That is, the probability of rejecting h_0 is greater when it is false than when it is true. Unbiasness, like the criterion of uniformly most powerful, is expressible using the β-function. A test is unbiased so long as $\beta(C, \theta_0) < \min_{\theta \in \Delta} \beta(C, \theta)$. A test is uniformly most powerful unbiased [UMPU], so long as its power function (β-function) is not exceeded (for each $\theta \in \Delta$) by any other unbiased test, i.e. it is most powerful among unbiased tests. As would be suspected, UMPU-tests are intimately related to C_{λ_Δ}-tests. In fact, A. Wald has shown that, for a broad family of distributions with independent trials, as the sample size increases indefinitely, likelihood ratio tests become asymptotically UMPU-tests.[28] Also, extending Fisher's results for UMP-tests, Neyman and Pearson have shown that (for single parameter problems satisfying several differentiability conditions on the probability function) a necessary condition for a UMPU-tests is the existence of a sufficient statistic.[29]

Where one-sided UMP-tests exist, but no UMP two-sided tests are available, it is often the case that an UMPU-test is constructable. For instance, reconsidering the survey sampling problem, and letting the family of alternatives

be two-sided, i.e. $\sim h_0$ including values of θ both greater than and less than $\frac{1}{2}$, we see that there is a UMPU-test (using the sufficient statistic $\bar{x}$) which is 'two-tailed.' $h_0: \theta = \frac{1}{2}$ is rejected whenever $\bar{x}$ falls too far from .5, i.e. whenever $|\bar{x} - .5| > \kappa$, where κ is a function of the size of the test. Also, this UMPU-test is a C_{λ_Δ}-test as well.

Finally, when neither UMP-tests nor UMPU-tests are present the N-P theory invokes a third condition, *invariance*, to rank possible tests. Invariance, unlike the other two constraints, is a symmetry condition which requires, loosely, that critical regions be invariant to alternative formulations of the sample space, seen as conventional selections, e.g. choice of origin and scale components in measurement, or orderings of the observations which are deemed arbitrary. As would be expected, the invariance condition carries only as much intuitive force as can be mustered in support of the specific claim that the characteristic defining the symmetry is a conventional element. In other words, conclusions of tests must be invariant over the special property used to delineate the symmetry. Optimally, invariance leads to a UMPU-I-test.

The N-P theory of hypothesis testing is completed by extending the conditions to cover composite null hypotheses. Apart from the obvious point that hypotheses can be composite in a multitude of ways, with corresponding classifications for comparable tests (especially restrictions involving size), no new criteria are introduced. Additional remarks on this topic must await the development of the Confidence Interval theory, and its ties to the theory of hypothesis testing.

2.5. HACKING'S CRITICISM OF N-P HYPOTHESIS TESTING

In his popular text, *Logic of Statistical Inference*, Ian Hacking argues that the N-P theory (sketched above) is inadequate for appraising the hypothesis being tested, once the data are available. He finds that, despite its reasonableness in terms of size and power constraints, as captured by the several criteria of UMP, UMPU and UMPU-I-tests, the plausibility disappears once the data are known precisely. That is, the investigator knows more than whether or not the outcome falls in the critical region (of even the best test) and this extra information can be relevant for determining the inductive support for the hypothesis tested.

Hacking promotes a theory of inductive support which is based on the likelihood function, given the data available. The overall compatibility of his program with the N-P theory goes as far as the mutual agreement that

C_λ and C_{λ_Δ} are the most desirable tests. It was described (above) how these likelihood ratio tests often survive as the N-P best tests; however, this agreement is not necessary, and Hacking gives an example (using a UMPU-I-test) for which the C_{λ_Δ}-test is biased, hence inferior. The force of his argument rests heavily on the intuitive superiority of likelihood ratio tests to any other. Since his counter-example involves only low level incompatibility with N-P, i.e. he invokes Invariance, the conflict might well be viewed as a family fight over what ought to be the lower level priorities. To repeat, the fundamental lemma for testing simple hypotheses against simple alternatives picks out C_λ-tests as uniquely best; and the necessary conditions for one-parameter UMPU-tests also involve C_{λ_Δ}-tests through the sufficient statistics used to define the critical region. Instead of following Hacking's strategy of attacking the N-P program at its hypothesis testing half, my criticism is reserved for the Confidence Interval theory; which is derivative on the testing program, and for which clear deficiencies exist using *UMP-tests*. That is, the faults Hacking claimed to be true of the hypothesis testing aspect of the N-P program can be clearly displayed in the confidence Interval portion, using only UMP criteria.

Hacking's example takes the following form:[30] the table below gives the distribution for the random variable x under the null hypothesis H and composite alternative $J \equiv (J_1 \text{ or } \ldots J_n)$.

	P(0)	P(n)	P(m) ($m \neq n$)
H	.90	.001	.001
J_n	.91	.090	0

The random variable can assume any of 101 values (0, . . . , 100). The probability for outcome $x = 0$ is .9, for H; and .91, for J. The probability for outcome $x = i\ (i \neq 0)$ is .001, for H; and .09, for J_i; and 0 otherwise.

The best likelihood ratio test of H against J, with size .1, uses a critical region of outcomes $x = i\ (i = 1, \ldots, 100)$. But this test is biased, with power equal to .09. Each outcome in the critical region has a low likelihood ratio: 1/90; but the overall region gives poor operating characteristic. On the other hand, the following 'mixed' test is a best unbiased-invariant test. Reject H iff 0 occurs and a randomizer yields a designated outcome with probability 1/9. This test rejects H on a subset of outcomes for which the likelihood test fails to reject, so H is never rejected by both tests. Hacking finds that the N-P program which favors the UMPU-I-test over the likelihood test does not

recognise the 'after trial' or 'backward looking' likelihood traits as best in all cases. Thus, the composite operating characteristic criteria of unbiased power and invariance are insufficient for determining satisfactory 'inductive' tests.[31]

Apart from the obvious point that Hacking's argument establishes only the incompatibility of UMPU-I-tests with likelihood tests, without thereby confirming the inferior status of the N-P best test, it is also the case that with increased sample size the likelihood test becomes recommended, even by N-P. That is, with repeated, independent samples from this same distribution, a likelihood ratio test (sometimes 'mixed,' because the distribution is discrete) has best operating characteristics.

Therefore, a serious advocate of the N-P program who felt obliged by Hacking's example to remove the conflict between likelihood tests and 'best' tests could either (a) suspend judgement in those cases where conflict exists, hoping for agreement with larger sample sizes, or (b) give up Invariance as a strongly held error condition whenever it leads to conflict with likelihood tests. This is not to down-grade the importance of small sample analysis. In fact, the Fisherian theory of estimation thrives on its answers to small sample problems. Nor is it intended to discredit Invariance as a legitimate criterion. Hacking's counter-example does undermine the conjecture that there is never a clash between good 'forward looking' tests and good 'backward looking' tests, assuming likelihood tests satisfy the latter requirement. It seems to me that the counter-example does not display the depths to which these two desiderata fail to agree. Therefore, it is my goal in part, to strengthen Hacking's evaluation by showing that N-P best tests lead to clearly inferior Confidence Intervals, based on UMP criteria alone. This is done in Section 2.7.

Before producing this evidence, it is worthwhile to review (briefly) another problem traceable to the forward look; specifically, the effect of the sampling rule on the selection of the rejection region.

2.6. THE SAMPLING RULE PROBLEM: ANOTHER CONSEQUENCE OF THE FORWARD LOOK

This chapter, dealing with the fundamentals of N-P hypothesis testing, would be left incomplete without a discussion of the so-called 'sampling rule paradox,' or the 'stopping rule paradox.'[32] A sample was said to be random if obtained by a method which makes it practically certain that, in the long run, the observed frequencies coincide with the probabilities of the stochastic model used to account for the phenomena under investigation.

For example, refer back to the elementary population survey problem

(p. 38). It was assumed that the sample of n individuals was a 'random' selection of the parent population with respect to F-ness. This was spelled out, in part, by appeal to a long run of 'hypothetical experiments,' under the assumption that the frequencies so produced would show that each individual in the parent population would be selected independently and equally often by rule Q. However, it was not stated whether rule Q stipulated that the sample be of size n, or whether the experimenter chose, for whatever reasons, to stop sampling after the nth member was observed.

Unfortunately, the sample space, i.e. the set of possible outcomes of the composite experiment, is not fully defined until this aspect of the problem is settled. For example, under one sampling rule, the one assumed for the survey problem, the sample size is predetermined; so that the sample proportion of F's, $\bar{x}$, is a random variable with [0, 1] range. Another potential sampling rule uses the $\bar{X}$ value to determine when to cease collecting evidence and perform the statistical test. For instance, rule Q' for the survey problem might require that, using the same device as rule Q, the sampling be terminated only when the $\bar{X}$ value falls within some narrow interval, e.g. when $\bar{x} \in [.5 \pm \delta]$. Clearly, the sample obtained during the experiment actually performed might satisfy both stopping rules. However, $\bar{x}$ is *not* a random variable under the FPS generated by rule Q'. Instead, it is n, the sample size which varies from one trial to another.

Therefore, the test for hypotheses about θ (the true proportion for F's in the parent population) uses one random variable ($\bar{x}$) under one sampling rule, and another (n) under the alternative stopping rule. What is most disturbing is that the best tests might not agree, given the same size (α level) and the same sample. The conflict is all the more puzzling when it is recognized that the likelihood, λ_Δ, for the data is the same no matter which stopping rule is actually adopted. Even the simple λ-likelihoods are the same.

By way of exposing the source of the possible conflict between the two best N-P tests, it must be remembered that C_{λ_Δ}-tests make use of other possible experimental outcomes, besides the one actually observed. Therefore, if (using one sampling rule) the observed outcome has a relatively low λ_Δ-value (compared to other values, based on different outcomes), then it will fall in the C_{λ_Δ}-region. However, if (using a modified sampling rule) that very same λ_Δ-value is relatively high (compared to other values based on the new set of possible alternative outcomes), then it will *not* fall in the C'_{λ_Δ}-region. Thus, fixing the likelihood does not guarantee equivalent rejection/non-rejection analyses over different stopping rules, using N-P test criteria. For the

orthodox statistical program, the intention of the investigator, as displayed in the stopping rule adopted, *is* relevant for the analysis.

This consequence of the N-P program has special application to the use of randomization in experimental design. The same sample can be the product of two, incompatible, randomizations. Each design (randomization) pattern has its own best test, and analogous to the foregoing case, these may not agree – even given the same data. There is little to say about the problem from the standpoint of the N-P theory: it is just another case of 'forward' versus 'backward' look difference. Prior to seeing the sample data, the two designs have different stochastic properties, with different sample spaces. Afterwards, these differences are retained, even though the actual data are an instance of either.[33]

2.7. CONFIDENCE INTERVALS

Without doubt, the theory of Confidence Intervals is the Neyman-Pearson substitute for a full-fledged program of inverse inference. As noted previously, using the FPS obtained from the model (which serves as the tie between the numerical measures of the FPS and the empirical frequencies generated by the 'repeated sampling' or 'randomness' claim), the unknown parameter cannot be construed as a random variable. That is, it remains an unknown constant so that the probability for: $\theta > k$ is either 0 or 1 (depending on whether the inequality is false or true). However, using the idea (based on the strategy of the example on p. 35) that probabilities for inequalities of the form $\theta \geqslant f(x)$ are available, where $f(x)$ is some known function of the random variable x (the data), we see that an alternative solution to the problem of inverse inference exists.

Perhaps the simplest method for explaining what a Confidence Interval is, is to rehearse one of Neyman's examples, used by him for the very same purpose.[34] Suppose that the data consist of two readings from independent trials of an experiment whose model is such that:

> the c.d.f. (cumulative distribution function) for the random variable x equals: 0, if $x \leqslant 0$
>
> $$x/\theta, \text{ if } 0 < x \leqslant \theta$$
>
> $$1, \text{ if } x > \theta;$$
>
> where it is known that $\theta > 0$.

That is, x is uniformly distributed between 0 and θ (θ positive); with density

function $\partial F/\partial x$:

$$0, \text{if } x \leqslant 0.$$

$$1/\theta, \text{if } 0 < x \leqslant \theta$$

$$0, \text{if } x > \theta.$$

The data (x_1, x_2), are two independent readings from this distribution. Using an elementary 2-dimensional sample space, we find that the density function, $\partial F/\partial(x_1, x_2)$ is uniformly $1/\theta^2$ for the square region $0 < x_1, x_2 \leqslant \theta$, and 0 everywhere else.

Consider, first, the random variable $x_L = \max(x_1, x_2)$, i.e. the greater of the two readings. It is easy to show that the probability distribution for x_L is given simply as:

$$\text{c.d.f. } F(x_L, \theta) \text{ equals: } \quad 0, \text{if } x_L \leqslant 0$$

$$x_L^2/\theta^2, \text{if } 0 < x_L \leqslant \theta$$

$$1, \text{if } x_L > \theta;$$

with density function $\partial F/\partial x_L$ equal to $2x_L/\theta^2$ in the positive sample space. Therefore, the probability for the inequality:

$$q \cdot \theta \leqslant x_L \leqslant \theta \qquad (\text{for } 0 \leqslant q \leqslant 1)$$

is just $1 - q^2$; independent of all other factors, including the unknown θ. Hence, with probability $1 - q^2$,

$$x_L \leqslant \theta \leqslant x_L/q. \qquad [\text{CI}_L]$$

This statement is to be distinguished from the formally similar

$$k_1 \leqslant \theta \leqslant k_1/q;$$

where k_1 is a constant; and for which only degenerate 0 or 1 probabilities are defined (depending on whether the inequalities are false or true).

The Confidence Interval arises by adopting the sequence of intervals for the unknown parameter defined by $[\text{CI}_L]$. It is to be remembered that the probabilistic ingredient of $[\text{CI}_L]$ stems from the FPS for the random variable, which has its connection to empirical frequencies only through the 'repeated sampling' property. Therefore, although the frequencies for the $[\text{CI}_L]$ statements are independent of the unknown parameter, hence remain 'accurate' even if θ changes from trial to trial, the force of any specific

interval (based on an observed data point) rests on the 'repeated sampling' properties of the continued use of $[CI_L]$, from trial to trial, qua *routine*.

To motivate the conceptual linkage between the confidence interval and hypothesis testing theories, note first that those intervals defined by $[CI_L]$ are not unique for their ability to produce Confidence Intervals for the unknown parameter. For example, let the random variable $\bar{x} = (x_1 + x_2)/2$ (the sample average) be used in place of x_L. It follows that $\bar{x}$ is distributed symmetrically about $\theta/2$, so that $2\bar{x}$ can be used as a central point for an interval about θ. In fact, the probability that $\theta - q' \leqslant 2\bar{x} \leqslant \theta + q'$ (assuming both x_i fall in the positive sample space) is just $1 - [(\theta - q')^2/\theta^2]$, $0 \leqslant q' \leqslant \theta$. Letting this probability value be α, we note the probability is α that:

$$\theta \cdot (1-\alpha)^{.5} \leqslant 2\bar{x} \leqslant \theta \cdot (2-(1-\alpha)^{.5}),$$

or

$$(2\bar{x})/(2-(1-\alpha)^{.5}) \leqslant \theta \leqslant (2\bar{x})/(1-\alpha)^{.5} \qquad [CI_{\bar{x}}].$$

Both $[CI_L]$ and $[CI_{\bar{x}}]$ generate confidence intervals (for any confidence level). No matter what the observed data, the intervals never agree: sometimes giving overlapping intervals, and sometimes such that the $[CI_L]$ interval is totally contained within the $[CI_{\bar{x}}]$ interval. However, the frequency basis for each set of intervals is exact, and built upon different random variables. Therefore, mere probability considerations cannot distinguish the superiority of one from the other, each being probabilistically valid.

Neyman's recommendation is to introduce a 'reliability' index to help rank the *informational* value of the several intervals available. An interval is considered more *reliable* or *informative* than another if it generates consistently narrower intervals. It is essential to remember that this factor is separate from the probability or frequency properties used to define the confidence interval concept. On this point Neyman says, (with changes for formulae names)

> If there is little to choose between the two systems of confidence intervals $[CI_L]$ and $[CI_{\bar{x}}]$, from the point of view of probability of correct statements, there are other aspects which easily determine the choice. In problems of estimation by interval, it is natural to try and get as narrow confidence intervals as possible. Comparing again $[CI_{\bar{x}}]$ and $[CI_L]$, we find that the latter interval is considerably shorter than the former. It is easy to see that this is a general rule. In fact, whatever the mean, $\bar{x}$, if both x_1 and x_2 are necessarily positive, then
>
> $$\bar{x} < L < 2\bar{x},$$
>
> . . . showing that the length of the confidence interval determined by $[CI_L]$ is always less than . . . that determined by $[CI_{\bar{x}}]$. It is obvious, therefore, that the confidence

intervals defined by $[CI_L]$ compared to the other system have definite advantages. These advantages, however, are independent of the conception of probability.[35]

The intuitive notion of using the length of a confidence interval to gauge its reliability is unsatisfactory, since it may be the case that, in comparing two different systems, for some values of the data one system yields the shorter interval while for alternative data the relation is reversed. Perhaps average or expected length could serve to salvage the intuition, but Neyman chooses a different route. He defines the system of 'shortest confidence intervals' as that for which the probability (at an assigned level) of 'covering' or 'including' values of the parameter other than the true one are a minimum, no matter what the true value be.[36] Thus, the pre-systematic desire for narrow intervals is converted into the precise requirement that the interval have low probability of covering false values of the unknown parameter.

This formulation also has the advantage that it quickly ties the confidence and hypothesis testing concepts together. A natural way for constructing confidence intervals is to consider a family of statistical tests for each of the simple hypotheses in the range of alternative values for the parameter, and then use the set of unrejected null hypotheses as the interval. The only (mild) constraint on this mode of construction is that the confidence interval should always be defined; so that it is impossible to rule out (reject) all possible values for the parameter, leaving an empty interval. Hence, confidence intervals can be formed by using the set of null hypotheses (h_0) for which the observed data fall in the 'acceptance region' under a test of that value of the parameter against the composite alternative of other values. Most remarkably, the criterion of 'shortest interval' (above) is then seen to be equivalent to the demand for UMP-tests as the basis or generator for the confidence interval.[37] That is, the intuitive demand for precise, reliable confidence intervals is translated back into the demand for best statistical tests, using the criteria of power and size, e.g. UMP-tests.

Continuing with the reduction of confidence interval criteria to hypothesis testing criteria, note that where UMP-tests are unavailable (which is a common enough phenomenon, when testing against a composite alternative), using unbiased-tests for devising confidence intervals permits the selection of 'unbiased intervals,' defined as: a set of intervals such that the probability of covering the true value of the parameter is greater than the probability of covering any false value, no matter what the true value be. Again, although weaker than the demand for 'shortest intervals,' unbiased intervals have an intuitive appeal, especially when a 'shortest, unbiased interval' can be constructed, based on a UMPU-test. Finally, by resorting to the condition of

Invariance of tests, 'shortest, unbiased invariant intervals' can be derived from UMP-U-I-tests. Of course, all the other properties of hypothesis testing, including the relation to likelihood tests (translated now as likelihood intervals) and the use of randomization for improved tests, spill over into the confidence interval theory, without modification.

Our account of Hacking's criticism is that the N-P theory of hypothesis testing is adequate for 'before trial' analysis, but fails for 'after trial' analysis, since the data are available for more detailed treatment than merely noting whether or not they fall into the critical region ('rejection region'). His counterargument relies on the fundamental superiority of likelihood tests, and the conflict between these and low level N-P criteria for best tests. It is claimed here that Hacking's conclusion (and general insight) is accurate but that more comprehensive evidence is available for its justification. The trick is to attack the N-P program through the Confidence Interval concept, and show it deficient in just the way Hacking claimed – only now using UMP-test based intervals.

Suppose the same stochastic model for the random variable x as in Neyman's example of p. 51, i.e. x uniformly distributed between 0 and θ (θ positive). For simplicity, assume only a single reading (although the conclusion below applies generally to any number of independent readings). It is easy enough to show that there is a UMP-test for the null hypothesis h_0: $\theta = \theta_0$, against the composite alternative $\sim h_0$: $\theta \in \Delta$ & $\theta \neq \theta_0$; where Δ is the set of alternatives defined by the model, i.e. $\theta > 0$. By the N-P lemma, this must be a C_{λ_Δ}-test, which is also a C_λ-test for any simple alternative, i.e. a likelihood ratio test. The likelihood for any alternative θ', to θ_0, for the data x is:

$$p(x/\theta_0)/p(x/\theta') = (1/\theta_0)/(1/\theta')$$

(assuming x in the positive sample space for each value of θ).

Therefore, the critical region for h_0 (at the α level) is made up of the union of x values for which the density (probability) is zero with those of minimum likelihood and such that the total probability is α. This is just the set of x: $(x \leqslant 0)$ or $(x > \theta_0)$ or $(x/\theta_0 \leqslant \kappa)$; where κ is a function of α, in fact $\kappa = \alpha$. This follows since the likelihood ratio is decreasing for decreasing θ', subject to the condition that $\theta' \geqslant x$.

Referring back to the results connecting UMP-tests to 'shortest intervals,' we find that the system of confidence intervals defined by:

$$\alpha\theta < x < \theta$$

or $\qquad x \leqslant \theta \leqslant x/\alpha \qquad [CI_\lambda]$

gives the best intervals (at the $(1-\alpha)$ probability level). $[CI_\lambda]$ is constructed from the set of null hypotheses *not* rejected by the UMP-test, described above.

For example, if the datum $x = 10$ is recorded, then the $[CI_\lambda]$ interval for the unknown parameter, at say the 95% level, is just [10, 200].[38]

Suppose now that the set of alternatives is modified so that it becomes known that $\theta \leqslant 15$. In other words, there exists an upper bound for the unknown parameter at the value 15. Since the tests of the previous page are UMP, restricting the set of alternatives by supplying an upper bound does *not* alter the UMP status. Therefore, the confidence intervals given by $[CI_\lambda]$ remain 'shortest,' even given the extra information about the upper bound. However, the plausibility of using the $[CI_\lambda]$, subject to the modified set of alternatives, is greatly reduced.

For instance, using the same datum as above, $x = 10$, we see that the new confidence interval is not [10, 200] (at the 95% level), but the truncated [10, 15] range, since it is known that $\theta \not> 15$. However, from the specification of the distribution, i.e. using just the model, it is known for sure that $\theta \geqslant x$. That is, it is known for certain that $\theta \geqslant 10$, once the observation is made. Also, given the upper bound, it is known that $x \leqslant \theta \leqslant 15$. But this is just the interval given by the rule $[CI_\lambda]$ at the 95% and not 100% level; all based on the 'shortest interval' and 'best test.' In other words, the most reliable interval, with the most information, tells nothing for the observation $x = 10$ (at the 95% level). If, in general, the upper bound is c_u, and the confidence level set at $(1-\alpha)$, then any observation in the range: $x \geqslant \alpha c_u$, leads to vacuously true intervals, i.e. true not merely at the $(1-\alpha)$ level, but at the 100% level. Thus, for the case above, with a 95% level, any observation: $x \geqslant \frac{3}{4}$, leads to a trivial confidence interval.

This consequence is totally independent of the frequency or 'repeated sampling' properties of the confidence intervals, even when subjected to the occasional deficiency, described above. That is, even when modified by the truncation, the $[CI_\lambda]$ intervals cover the true value (in the long run) exactly $(1-\alpha)$% of the time; even if the unknown parameter varies from trial to trial (in any manner). What has happened is that more information is available (in some cases) than merely that the $[CI_\lambda]$ rule has been applied to the data obtained. To use Hacking's formulation, it seems reasonable to say, *before knowing the data*, that the $[CI_\lambda]$ interval will cover the true value of the unknown parameter with probability $(1-\alpha)$; however, *having seen the value* x, it may be unreasonable to maintain the probability statement or use it

to express a degree of confience in the interval generated by the $[CI_\lambda]$ rule. This is so, despite the selection of $[CI_\lambda]$ as best according to the very highest N-P criteria.[39]

The failure of the N-P 'best' test to produce 'shortest' confidence intervals with wholly desirable properties on the backward look is emphasized even more by considering the following alternative test, with derivative intervals, having the properties that:

a – when the parameter space is unbounded from above the intervals are of infinite length, for all sample points;

b – when the parameter space is bounded from above the truncated intervals are generally incompatible with those from the 'best' test; however, the set of sample points generating logically trivial intervals is a proper subset of those with the same undesirable property for the 'shortest' interval;

c – the test underlying the new intervals is biased, both for the bounded and unbounded cases.

Hence, for the truncated version stemming from the addition of an upper bound on the unknown parameter, the confidence intervals generated by a biased test can display superior qualities on the backward look: having the degenerate interval property on only a proper subset of the data points leading to the same, useless results for the 'shortest' interval. Of course, the 'shortest' intervals are clearly superior for the original, unbounded case.

The alternative test is described as follows. Unlike the best test, which is UMP, define a 'central' critical region $C_{alt.}$, at the α level, $(\alpha \leqslant .9)$ as:

$$C_{alt.} = \{x: (1 - [1.1 \cdot \alpha])\,\theta \leqslant x \leqslant (1 - [.1 \cdot \alpha])\theta\} \cup \{x: x < 0 \text{ or } x > \theta\}.$$

This leads to a (sometimes disconnected) confidence interval, at the $(1-\alpha)\%$ level, defined as:

$$x \leqslant \theta \leqslant x/(1 - [.1 \cdot \alpha]) \quad \text{and} \quad x/(1 - [1.1 \cdot \alpha]) \leqslant \theta. \qquad [CI_{alt.}]$$

For the unbounded cases, the second component yields infinite intervals only. Therefore, without the upper bound, the $[CI_\lambda]$ are superior. However, if it becomes known that $\theta \leqslant 15$, the truncated $[CI_{alt.}]$ are of finite length. Without any difficult calculation it is clear that the $[CI_{alt.}]$ intervals are logically trivial only for data such that: $x \geqslant (1 - [.1 \cdot \alpha])u$, where u is the upper bound for θ. Comparing this to the range of data points for which the 'shortest' interval yields trivial results at the 95% level, and using $u = 15$, $[CI_{alt.}]$ is degenerate for $x \geqslant 14.925$, while $[CI_\lambda]$ is degenerate for all

$x > \frac{3}{4}$. Thus, the inferior confidence interval has the undesirable property for only a small region of those outcomes for which the 'most informative' interval is useless.

Finally, the $C_{\text{alt.}}$-test is strongly biased for alternatives θ' such that $\theta' \leqslant (1 - [1.1 \cdot \alpha])\theta_0$; where θ_0 is the null hypothesis of the test. This follows since, for θ' thus defined, the power of the test drops to 0, since the critical region $C_{\text{alt.}}$ lies wholly above θ', i.e. in a region of zero density. This is not affected by the imposition of an *upper* bound, because the bias remains for alternatives *below* the parametric value specified by the null hypothesis.

In conclusion, it is claimed that the N-P theory fails on the 'after trial' analysis since the 'best' interval, based on the 'best' test can be logically trivial (for almost all of the sample space); whereas a N-P *inferior* test and *inferior* interval has superior 'after-trial' consequences.

2.8. COMPOSITE CONFIDENCE INTERVALS AND CONFIDENCE EQUIVALENCE CONDITION

Although the foregoing discussion, using one parameter distributions, served to illustrate the tie between the theory of hypothesis testing and confidence intervals, a more realistic setting includes multiparameter distributions and their more complex family of related problems. These situations can be approached in either of two ways. First, the underlying statistical tests can be simple, i.e. can specify exact values for all parameters, generating 'simultaneous' confidence intervals.[40] For example, if the unknown parameter $\theta = (\theta_1, \theta_2)$ is two-dimensional, then simultaneous confidence intervals can be built up from the family of hypothesis tests $h_0 : (\theta_1 = k_1)\ \&\ (\theta_2 = k_2)$; where k_1 and k_2 are specified constants. As before, the confidence interval is made up of all un-rejected null hypotheses.

Alternatively, 'composite' confidence intervals can be analyzed, where the underlying statistical tests are composite. These are preferred to 'simultaneous' intervals because they are the multivariate analogues of the corresponding one parameter problems. For example, if (as above) $\theta = (\theta_1, \theta_2)$, then the simple hypothesis $h_0 : \theta_1 = k_1$, when θ_2 is known, becomes a composite hypothesis when θ_2 is not known. However, reliance on tests of composite null hypotheses to generate confidence intervals severely restricts the program, since these are seldom of the sort to allow for *exact* confidence intervals. That is, composite null hypotheses usually admit only test whose size (α level) is defined by an inequality. Therefore, the derivative confidence intervals have merely the (weakened) property of 'covering' the true value of

the unknown parameter with a probability *not less than* $(1-\alpha)$, where α is the size of the generating statistical test.

When, however rare, exact statistical tests for composite null hypotheses exist, an interesting condition can be imposed on the 'composite' confidence intervals which they produce. Two systems of confidence intervals are said to obey the novel rule of *Confidence Equivalence* just in case they cover (fail to cover) the same parametric value, when the hypotheses used to generate the intervals are logically equivalent. For example, the composite hypothesis $h_0: \theta_1 = 0$ is logically equivalent to $h_0': (\theta_1/\theta_2) = 0$. Therefore, a system of confidence intervals for hypotheses about θ_1 and a system of confidence intervals for hypotheses about the ratios (θ_1/θ_2) satisfy Confidence Equivalence [CE] just in case both include (fail to include) the value 0, for common data. Remembering the foundational tie to statistical tests, we see [CE] requires that the two null hypotheses h_0 and h_0' be treated similarly, i.e. both rejected or both unrejected, by their corresponding tests for common data.

Confidence Equivalence, although a plausible condition, represents a criterion different in kind from the other concerns, e.g. power and bias. The operating characteristics focused on 'long run' consequences of repeated use of the test rule. The counterpart restraints on confidence intervals considered only long term dividends of using the confidence rule, e.g. 'shortest' or 'shortest-unbiased' intervals. CE, on the other hand, imposes a consistency requirement which must be met at each stage in the sequence of repeated trials. However, in the spirit of the other N-P concerns, the CE rule is 'forward looking,' since it builds in ahead of time a result which is displayed also in the long run, as well as in the short run. Intuitively, CE demands that the logical formulation of the question not affect the rejection/non-rejection status of the answer. Perhaps, with only mild strain, it can be considered an extension of the Invraiance condition; the symmetry being equivalence over 'logical transformation' of the parameter(s).

It is important to see that CE does not necessarily compete with the other conditions, and that it can serve as the basis of a defense for less intuitively desirable consequences, e.g. vacuous or logically trivial intervals. In the example below, not only will these claims be defended but a simple methodological procedure for deriving CE systems will be supplied. It is called the Master Significance Test [MST].

Two practically important and historically significant distributions are the normal and bivariate normal. The full bivariate normal distribution for the two-dimensional random variable (x, y) involves five parameters: (μ_x, μ_y) – the two population means, (σ_x^2, σ_y^2) – the two population

variances, and ρ_{xy} – the population correlation coefficient. It will be assumed that composite hypotheses involve the unknown means only, with the other three parameters known. For simplicity, numerical examples will be given assuming $\sigma_x^2 = \sigma_y^2 = 1$ and $\rho_{xy} = 0$: a bivariate normal with unit standard deviation and independence between the two random variables. However, the important formulae are given in full generality.

It is well known that the best tests for the hypotheses about the two unknown means (separately) make use of regions defined in terms of the respective random variable alone. That is, the confidence intervals for $\mu_i\ (i = x, y)$ are given by:

$$|\mu_i - i| \leqslant \kappa$$

or

$$i - \kappa \leqslant \mu_i \leqslant i + \kappa; \qquad [\mathrm{CI}_{\mu_i}]$$

where κ is a function of the confidence level. Thus, the two sets of intervals for the 'composite' hypotheses about the two means are actually uniparameteric, since the other random variable (y for μ_x and x for μ_y) plays no role. That is, the underlying statistical test is actually a UMPU-test, based on the single sufficient statistic $i\ (i = x, y)$.

A similar situation arises for tests of certain composite 'linear' hypotheses, e.g. $\xi = (\mu_x + \mu_y)$ and $\zeta = (\mu_x - \mu_y)$, for which the best tests are:

and

$$|\xi - (x + y)| \leqslant \kappa' \qquad [\mathrm{CI}_\xi]$$

$$|\zeta - (x - y)| \leqslant \kappa'. \qquad [\mathrm{CI}_\zeta]$$

The recommended confidence interval for hypotheses about ratios, $\gamma = (\mu_x/\mu_y)$, given by E.C. Fieller,[41] 'Some problems in interval estimation,' is defined by:

$$|y| \cdot (\gamma - [x/y])/\sqrt{1 + \gamma^2} \leqslant \kappa. \qquad [\mathrm{CI}_\gamma]$$

However, the Fieller interval suffers from the defect that, given a confidence level, there exist data points (x, y) such that $[\mathrm{CI}_\gamma]$ yields only trivial intervals, i.e. $\gamma \in (-\infty, +\infty)$. And, for any datum point, there exists a confidence level such that (at that level or greater) the $[\mathrm{CI}_\gamma]$ interval is logically vacuous.

The mathematical details for this case are given in an appendix, however, an intuitive grasp of the problem can be had through the following picture-description. In the two-dimensional (μ_x, μ_y) plane, hypotheses about the ratio γ (or γ^{-1}) are represented by straight lines through the origin (0, 0), with different slopes. Superimposing the (x, y) sample space onto the para-

meter space, sample points with a constant likelihood ratio for a null hypothesis are located on straight lines parallel to the null hypothesis and at a given distance (on either side) from it. Acceptance regions are, therefore, belts of constant width encompassing the null hypotheses not rejected by this test. This amounts to rejecting all null hypotheses not falling within a circle of radius r from the sample point; where r is a function of the confidence level. Referring back to the underlying test, we see that the width of the acceptance region increases with decreasing size; hence, the radius of the circle of acceptance increases with increasing confidence level. As the sample point moves towards the origin, or as the confidence level increases, the circle shifts to include the parametric point (0, 0). Once this happens no null hypothesis is rejected by the test, and the $[CL_\gamma]$ interval is vacuous. This is because all ratios 'go through' the origin; which is a singularity point for the partition of alternative ratio hypotheses.

This deficiency of the Fieller test caused some concern and was the motivating force behind an alternative proposal, made by M. Creasy,[42] 'Limits for the ratio of means,' and which avoided the Fieller pitfall. The Creasy solution is also a Fiducial probability for the ratio problem, and will be analyzed with respect to that concept in a later section. By N-P criteria, her answer yields only inferior confidence intervals, and Fieller's $[CI_\gamma]$ are preferred.

Besides operating characteristic considerations, CE enters in support of Fieller's solution. That this is so is evident from the following argument, which supplies yet another derivation of the standard confidence intervals for the foregoing bivariate normal problems.

From certain simple and well known properties of the bivariate normal distribution[43] it can be shown that the MST is a standard normal variate, with 0 mean and unit variance. The MST (Master Significance Test) is defined as:

$$[\text{MST}] \quad \frac{a(\mu_x - x) + b(\mu_y - y)}{[a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\sigma_{xy}]^{.5}} \cdot \quad (44)$$

By specifying values for a and b, families of homogeneous linear hypotheses about the unknown means are fixed. The rejection rule for the MST is to reject values for the two unknown means (given the values of the two variances, the correlation, a and b, and the sample data x and y) whenever the absolute value for MST exceeds a known constant (whose value is determined by the confidence level).

For example, setting:

$b =$ 0 yields hypotheses about μ_x (marginal)

$a =$ 0 yields hypotheses about μ_y (marginal)

$0 \neq a =$ b yields hypotheses about $\mu_x + \mu_y$ (sums)

$0 < a = -b$ yields hypotheses about $\mu_x - \mu_y$ (differences)

$-1 = a, \quad b = \mu_x/\mu_y$ yields hypotheses about μ_x/μ_y (ratios)

A 95% confidence interval is generated from a statistical test of size (α level) .05; which for a standardized normal variate uses a 'two tailed' critical region of values more than 1.96 (standard deviation) units from the mean (0, for the MST). Hence, the 95% confidence interval generated by the MST consists of all values for the unknown means μ_x and μ_y which keep the MST value within $\pm$ 1.96 of zero. For example, the MST confidence interval for ratios $\mu_x/\mu_y = \gamma$, using $a = -1, b = \gamma$, yields:

$$(\gamma \cdot -y + x)/[1 + \gamma^2]^{\cdot 5} \in [-1.96, +1.96]$$

(for the special values: $\sigma_x^2 = \sigma_y^2 = 1$ and $\rho_{xy} = 0$). This agrees, exactly, with the Fieller solution.

The special property of MST intervals is that they obey CE. That this is so is clear once it is recognized that the logical equivalence for certain hypotheses, at particular parametric values, is reflected in the MST through the a, b values. For example, that $\mu_x = 0$ iff $\gamma = 0$ is translated into the proposition that (with $b = 0$ for testing hypotheses about μ_x) the MST interval covers the value $\mu_x = 0$ just in case (with $a = -1$, for testing hypotheses about γ) $b = 0$. But this is just the equivalence needed for CE. Therefore, even though the Fieller solution (and the MST solution) for confidence intervals about ratios have the unfortunate property of sometimes generating logically vacuous ranges, they have the important beneficial property of being consistent with the intervals from other, logically related, questions.[45]

The general technique for producing Master Significance Tests relies on a family of transformations which keeps some hypothetical random variable within the same distribution model. For the bivariate normal, the set of transformations is as large as the family of linear transformations of the two random variables. This includes all composite linear hypotheses for the two unknown means, as well as the degenerate (marginal) transformations for the two parameters separately. With this transformation group (and it is a mathematical group), the MST can be extended to a tri-variate normal

by including a third component, c, for the third random variable, z, and its unknown parameter, μ_z. In this case, the bivariate model can be recovered by seeing the MST as a special case with $c = 0$. Unfortunately, however, the set of questions covered by this addition is not a simple extension of the two-dimensional ones, e.g. one cannot use the extended test to answer whether $\mu_x = \mu_y = \mu_z$, whether the means are equal. For two parameters the question is answered, using the MST, by determining whether $(\mu_x - \mu_y = 0)$ is rejected.

This concludes the discussion of the foundations of Neyman-Pearson Hypothesis Testing and Confidence Interval Theory. Our conclusions are that the two programs are mutually compatible, that they fail the 'after trial' evaluation, and finally that the augmentation to include CE mitigates only a few of the serious deficiencies.

Before turning to the discussion of Fisher's alternatives to the Neyman-Pearson solutions of inverse inference, the following example vividly displays many of the features already discussed in this chapter. Also, the example involves a vexing statistical problem, known as the Behrens-Fisher problem.

2.9. AN EXAMPLE ILLUSTRATING THE USE OF RANDOMIZATION FOR CONSTRUCTING CONFIDENCE INTERVALS

One of the more famous puzzles for which no exact N-P confidence solution exists is the so-called Behrens-Fisher (bivariate normal) problem about the difference in means $(\mu_x - \mu_y)$, when the variance σ_x^2 and σ_y^2 are unknown, and the correlation is known to be 0. A full discussion of the Behrens-Fisher solution is given in a later section, dealing with multivariate fiducial probability and the step-by-step process. The surprise surrounding this problem stems, in part, from the statistical fact that for the normal distribution confidence intervals exist for hypotheses about the unknown mean, even when the variance is unknown. These are given by a 'Studentized' version of the MST test, where the population variance σ^2 is replaced by an estimated variance s^2, based on the sample data.[46] Moreover, with the correlation, ρ_{xy}, known to be zero, the random variables are distributed independently, so that it would seem intuitively plausible that a N-P confidence interval for the difference in population means might be constructed from the two confidence intervals for the means taken separately. However, when this strategy is employed the resulting distribution (the Behrens-Fisher distribution) is seen to be a function of the unknown variances, the variance ratio σ_x^2/σ_y^2, in fact. This nuisance factor enters to the disadvantage of the proposed N-P solution.

No exact confidence level is produced by this technique, since the underlying test does not have exact size, but merely an inequality (based on the unknown variance ratio).[47]

An alternative N-P solution (allegedly suggested by Neyman) with exact frequencies is available; but it requires the use of a randomizer.[48] The key to Neyman's solution is that the difference between two independently distributed normal variates is again normally distributed. More importantly, the mean of this third distribution is just the difference in the means of the two original distributions; which is the parametric value under inquiry. Finally, the variance for this third normal distribution is the sum of the two original variances. Therefore, using a sequence of differences between the sample data, defined as $z_i = x_i - y_i$, a Student's test for hypotheses about $\mu_z = \mu_x - \mu_y$ is made available; where $\bar{z}$ (the z average) replaces the sample means, and s_z^2 (the z sample variance) replaces the two x, y sample variances.

The only wrinkle in this line of argument is that the test retains its validity so long as the two samples are paired in such a manner that the correlation remains at its initial value, i.e. 0. To insure this, a randomizer is employed to form the z sequence by selecting arbitrary x and y values for pairing. Technically, the difference in pairing procedures is reflected in the s_z^2, the sample variance for the z sequences. That is, no matter how the samples are paired, $\bar{z} = \bar{x} - \bar{y}$. But the z estimate of variance fluctuates from one pairing pattern to another. The smaller the s_z^2 value produced, the greater the emphasis on the observed $\bar{z}$. In other words, Neyman's solution gives narrow intervals about $\bar{z}$, when the pairing process yields a small s_z^2, and a broad interval otherwise.

Using the 'before trial' versus 'after trial' analysis, it is the case that the Neyman solution yields an interval, with the appropriate frequencies, before seeing the selection determined by the randomizer. However, once the pairing is done, it is possible to compare the s_z^2 generated against those which might have been produced (by the randomization), to see just how much influence the pairing process had. For example, it might be the case that the observed s_z^2 is the theoretical minimum for the data collected. It does not seem reasonable to claim great precision, in such a case, just because of the fortuitous random pairing.

NOTES

[1] This remark is hardly controversial. Another writer (Kyburg) goes so far as to baptize it the 'classical statistical theory.' H.E. Kyburg, Jr., *The Logical Foundations of Statistical Inference*, (Dordrecht and Boston: D. Reidel, 1974), p. 22. More traditionally, it

is referred to as the 'orthodox' or 'British-American School' theory. See, for example, Lindley, *Introduction to Probability and Statistics* or Leonard J. Savage, *The Foundations of Statistics*, 2nd ed. (New York: Dover Publications, 1972).

[2] The two early papers: J. Neyman and E.S. Pearson, 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference', Part I, *Biometrika* 20A (1928), 175–240. Rep. in *JSP*, paper #1. Part II, *Biometrika* 20A (1928), 263–94. Rep. in *JSP*, paper #2. For Neyman's theory of confidence intervals, see: J. Neyman, 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', *Phil. Trans. Royal Soc. of London*, Ser. A 236, #767 (1937), 333–80. Rep. in *SESPJN*, paper #20. Also, the reference for footnote 3, chapter 1, *SESPJN*, paper #10, p. 121.

[3] A complete presentation of Wald's views is found in: Abraham Wald, *Statistical Decision Functions*, 2nd ed. (New York: Chelsea Publishing, 1971). For an earlier, preliminary version emphasizing the connection to N-P theory, see: A. Wald, *On the Principles of Statistical Inference* (Notre Dame: University of Notre Dame, 1942).

[4] A summary of Robbins' ideas is found in: Herbert Robbins, 'A New Approach to a Classical Statistical Decision Problem', in H. Kyburg and E. Nagel ed., *Induction: Some Current Issues* (Wesleyan University Press, 1963), pp. 101–113.

[5] Similar criticism is given by A.P. Dempster in 'On Direct Probabilities', *J. Royal Stat. Soc.*, B 25 (1963), 100–110. Also, in his 'On the Difficulties Inherent in Fisher's Fiducial Argument', *J. Am. Stat. Ass.* 59 (1964), 56–66.

[6] H. Reichenbach, *The Theory of Probability*, p. 374. Kyburg's program, Epistemological Probability, discussed in Chapter 7, improves Reichenbach's vague prescription for selecting the reference class.

[7] Obviously, the 'narrowest reference class' does not always generate statistically useful classes. It might very well be the case that although two reference classes have known statistics, i.e. limits or frequencies of whatever serves as the explication of the probability$_2$ term, their intersection may generate a reference class with *unknown* statistics. Therefore, Reichenbach's rule is qualified by the condition that the narrowest reference class with known (or *epistemically* precise) statistics be used. This leads to the somewhat unsatisfactory conclusion that a reference class may be selected for its epistemically precise statistics, even though it is *known* that a particular sub-class has *different* statistics. What saves the larger reference class is that the different statistics for the narrower reference class are only known *imprecisely*. An extreme version is where the sub-class is known to have different frequencies, but are otherwise unrestricted in the [0, 1] range.

[8] J. Neyman, 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', rep. in *SESPJN*, p. 253.

[9] *Ibid.*, p. 255.

[10] The law of big numbers to which Neyman refers, unfortunately, is merely a high probability$_2$ law. That is, a theorem whose statistical conclusion is an outcome with probability$_2$ $1-\epsilon$. Assuming just normal background evidence, a direct inference is available for a particular sequence of 'repeated trials' (say, the *next* run of 10^{10} trials). Hence, this high probability$_2$ can be converted to a high probability$_1$ about the frequency of outcomes in the long run under consideration. However, even a high probability$_1$ does not guarantee the conclusion of the direct inference; to wit, that the frequencies will remain within a narrow range, $[\kappa \pm \delta]$. Therefore, Neyman's attempt

at explaining the application of the probability theory in terms of 'random samples' by appeal to the law of big numbers does *not* eliminate probability, in favor of *actual* frequencies generated by repeated sampling. By the direct inference, let the conclusion have the force of a 'practical certainty,' where the certainty applies to the epistemological state of the investigator and *not* to the modality of the empirical facts, i.e. outcomes of the repeated trials. In other words, the investigator is practically certain that, on repeated trials, the observed frequencies correspond to the mathematical measures over the FPS.

[11] J. Neyman, *op. cit.*, p. 255.

[12] *Ibid.*, p. 256.

[13] *Ibid.*, p. 257.

[14] Hacking, *Logic of Statistical Inference*, pp. 89–114.

[15] *Ibid.*, p. 93.

[16] *Ibid.*, p. 98. This example is discussed in the text, p. 48.

[17] J. Neyman, *First course in Probability and Statistics* (New York: Henry Holt and Company, 1950), §5.1.2.

[18] *Ibid.*, §5.4, definition 5.4.

[19] The conditions under which these frequencies are exact are discussed below.

[20] J. Neyman and E.S. Pearson, 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Phil. Trans. Royal Soc., Ser. A* 231 (1933), 289–337. Rep. in *JSP*, paper #6, p. 144.

[21] This is known as the Fundamental Neyman-Pearson Lemma. See, *ibid.*, *JSP*, p. 149. Without continuity of the random variable, as in the example above, the most powerful test is a *mixed* likelihood ratio test. Then a randomizer is used to 'smooth out' the distribution, as in the first example of the next section.

[22] With discrete distributions likelihood ratio tests will not be defined for every size, i.e. for every value of α. In such cases one may use the likelihood ratio test of largest size not exceeding the required level, or a *mixed* likelihood ratio test to achieve the exact α level demanded (as in the first example of the next section). This problem carries over to the testing of simple hypotheses with composite alternatives, discussed in Section 2.4.

[23] Kyburg, *The Logical Foundations of Statistical Inference*, p. 52. The Cox paper is: D.R. Cox, 'Some Problems Connected with Statistical Inference', *Annals of Mathematical Statistics* 29 (1958), 357–63. Kyburg uses the example as a *reductio ad absurdum* of the operating characteristic approach.

[24] J. Neyman and E.S. Pearson, 'The Testing of Statistical Hypotheses in Relation to Probabilities A Priori', *Proc. Cambridge Phil. Soc.* 24 (1933), 492–510. Rep. in *JSP*, paper #7, p. 192.

[25] Although, in general, sufficient conditions for the existence of UMP-tests are not available (short of the defining conditions themselves), an important necessary condition based on the N-P lemma exists. In his historic paper, 'Two New Properties of Mathematical Likelihood', §2.6, R.A. Fisher argues that UMP-tests exist only if: (i) the class of hypotheses admits of a one-parameter formulation, and (ii) there exists a sufficient statistic for the parameter involved. [R.A. Fisher, 'Two New Properties of Mathematical Likelihood,' pp. 294–295.] A thumbnail sketch of his proof for these claims goes like this: assuming a UMP-test exists, it must be a C_{λ_Δ}-test such that all values of the random variable of constant likelihood on one simple alternative are of constant likelihood on any other simple alternative. Moreover, if T' is the statistic defining these 'contours of

constant likelihood,' then the likelihood for any alternative must be expressible as $\kappa \cdot p(T'/\theta)$; where θ represents the unknown parameter(s), and κ is a function independent of θ. This will happen, argues Fisher, only if T' is a function of one parameter, i.e. θ can be replaced by a single parameter, and also if the boundary of the critical region defined in terms of T' (which happens to be the maximum likelihood estimate for the parameter) has a distribution independent of the parameter. Thus, T' is also sufficient for the unknown parameter.

The factorization condition $\kappa \cdot p(T'/\theta)$ for the likelihood function can be connected directly to the sufficiency of T' for θ by use of Neyman's Theorem for necessary and sufficient conditions of a sufficient statistic. His proof dates from 1935. [J. Neyman, 'Su un teorema concernente le cosiddette statistiche sufficienti,' *Giornale dell'Istituto Italiano degli Attuari* VI, #4 (1935).] However, Fisher's paper (cited above) antedates it by over a year, so the more direct proof was not yet available.

In 1936 Neyman and Pearson responded to the earlier Fisher paper, rebutting his claims with the denial of both. [J. Neyman and E.S. Pearson, 'Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses,' in *JSP*, paper #9, pp. 248–251 and pp. 254–264.] With respect to Fisher's result (i), it must be admitted that the restriction to uni-parameter problems is unnecessary, as demonstrated in the N-P paper §5. However, apart from a minor correction (concerning the dimensionality of the boundary for the critical region), the second result (ii) is correct. More accurately, the existence of a UMP-test entails that the portion of the *positive* sample space, given h_0, i.e. that region of the sample space with probability greater than zero given h_0, which falls in the critical region (at any size) must be specifiable using a single 'sufficient' statistic: which is truly sufficient, given the additional information that it falls in the positive region of the sample space under h_0.

[26] J. Neyman, 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability,' rep. in *SESPJN*, p. 283.

[27] J. Neyman and E.S. Pearson, 'Contributions to the Theory of Testing Statistical Hypotheses', *Statistical Research Memoirs* 1 (1936), 1–37. Rep. in *JSP*, paper #8, p. 210.

[28] Abraham Wald, 'Asymptotically Most Powerful Tests of Statistical Hypotheses', *The Annals of Math. Stat.* XII, #1 (1941), 1–19. Rep. in *Selected Papers in Statistics and Probability by Abraham Wald* (Stanford: Stanford University Press, 1957), *SPAW*.

[29] J. Neyman and E.S. Pearson, 'Contributions to the Theory of Testing Statistical Hypotheses', Part II, *Statistical Research Memoirs* 2 (1937), 25–57. Rep. in *JSP*, paper #10, p. 274.

[30] Ian Hacking, *Logic of Statistical Inference*, p. 98. Hacking uses P to denote probability$_2$.

[31] *Ibid.*, pp. 99–102.

[32] A brief discussion is given by Leonard Savage, 'The Foundations of Statistics Reconsidered,' rep. in H. Kyburg and H. Smokler eds., *Studies in Subjective Probability*, pp. 184–5.

[33] The final chapter, dealing with Fisherian policies for experimental design (which are intended to reflect both the 'forward' and 'backward' look), contains a brief discussion of this last issue, along with a tentative solution.

[34] J. Neyman, 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability,' in *SESPJN*, p. 271. Neyman's analysis for subcase 2 is inaccurate,

as discovered by Kyburg. The text gives the corrected formulas.

[35] *Ibid.*, p. 274.

[36] *Ibid.*, p. 282. It is to be noted that S.S. Wilks has investigated the properties of 'shortest' confidence intervals using the criterion of shortest expected length: S.S. Wilks, 'Shortest Average Confidence Intervals from Large Samples', *Annals of Math. Stat.* 9 (1938), 166–75. Also, A. Wald, 'Asymtotically Shortest Confidence Intervals', *Annals of Math. Stat.* XIII, #2 (1942), 127–137. Rep. in *SPAW*, pp. 207–17, especially pp. 216–7. Wald shows conditions under which there is asymptotic agreement between the two alternative approaches to 'shortest' intervals.

The example in the text, pp. 51–2 (taken from Neyman) does *not* meet the conditions required for Wald's analysis to apply. The positive sample space depends upon θ, hence the distribution is not 'regular.' For results that do apply see J.K. Ghosh, 'On the Relation Among Shortest Confidence Intervals of Different Types', *Calcutta Stat. Assoc. Bul.* 10 (1961), 147–152. Also, see J.W. Pratt, 'Length of Confidence Intervals', *J.A.S.A.* 56, #295 (1961), 549–567.

[37] J. Neyman, *op. cit.*, p. 282. It is to be noted that Neyman's 'shortest' intervals are also known as 'most accurate' confidence intervals, to distinguish them from Wilks' intervals.

[38] For the general case of a sample of n independent readings there is a UMP-test based on the sufficient statistic x_L, defined as $x_L = \max(x_1, \ldots, x_n)$. The $[CI_L]$ interval, defined on p. 52, is the special $n = 2$ case of this family. It follows that the $[CI_L]$ intervals are the 'shortest' in general.

[39] A similar phenomenon results in an example, due to B.L. Welch, and discussed on p. 152. The source of the difficulty is modified knowledge of the unknown parameter based, in part, on the value of an ancillary statistic (which in and of itself tells nothing about the parameter). The ancillary statistic does serve, as the randomizer of the example on p. 44, to knit together tests of different precision. That is, the value of the ancillary statistic specifies the precision of the best test, given the rest of the data. However, sometimes (when the ancillary statistic shows the test could be made very precise) trivial intervals are generated so as to allow for a more informative interval when only less precise tests are available. In other words, there is a trade-off such that in some cases vacuously true intervals are generated, whereas in other cases a narrow (informative) one can be used, all while keeping the overall long run frequencies at the proper values.

Further clarification is available by reviewing the randomized most powerful test of p. 44. A trivial (vacuously true) interval arises when the underlying statistical test fails to reject any null hypothesis. When the randomizer (ancillary statistic) yields the more precise experiment ($\sigma^2 = 1$), the size drops to a value very nearly equal to 0, i.e. it is practically certain that the null hypothesis will *not* be rejected when $\sigma^2 = 1$. This is analogous to the situation in the Welch example.

[40] The construction of multiparameter confidence intervals is not standardized. For example, H. Cramer, *Mathematical Methods of Statistics* (Princeton: Princeton University Press, 1971), pp. 516–17, uses what the text calls 'simultaneous' intervals; whereas D.A.S. Fraser, *Statistics: an Introduction* (New York: John Wiley and Sons, 1958), pp. 284–7, uses what the text calls 'composite' intervals for 'Multiple Confidence Methods.'

[41] E.C. Fieller, 'Some Problems in Interval Estimation', Symposium on Interval

Estimation, *Royal Stat. Soc., Ser. B* 16 (1954), 175–185.

[42] M. Creasy, 'Limits for the Ratio of Means,' Symposium on Interval Estimation, *Royal Stat. Soc., Ser. B* 16 (1954), 186–94.

[43] See appendix #9.7.

[44] The test is called a 'significance' test due to the character of the rejection rule. It is more in the spirit of a Fisherian test (specifying just a rejection region, without any power considerations) than a N-P hypothesis test.

The Master Significance Test is, to the best of my knowledge, novel for constructing families of related confidence intervals. However, it is to be noted that Fisher's comments in the Discussion on the Symposium on Interval Estimation, pp. 212–13, served as the stimulus for the test.

[45] Fieller recognized some of these consequences (for his own test) relating to the linear regression theory for the bivariate normal distribution. See, especially, his remarks in his Appendix A.2, pp. 184–5. The stated agreement of the MST with his solution depends upon the symmetry of normal distribution about the mean.

[46] A full discussion of this is presented in a later section, pp. 141–146.

[47] This result is well known. For a concise summary of the problem for confidence interval theory see: J. Neyman, 'Fiducial Argument and the Theory of Confidence Intervals', *Biometrika* 32 (1941), 128–50. Rep. in *SESPJN*, paper #26, pp. 383–7.

[48] Neyman's solution is reported in R.A. Fisher, *Statistical Methods and Scientific Inference*, pp. 101–3.

CHAPTER 3

FISHERIAN SIGNIFICANCE TESTING

Having reconsidered and rejected the orthodox position on inverse inference, I turn to the major task, reconstructing Fisherian inference. Section 3.1 provides a preliminary discussion of Fisher's notion of an *hypothetically infinite* reference class and its connection to chance (probability$_2$) and direct inference. A brief analysis sketches out the similarity between Fisher's and Kyburg's views on probability, especially the importance both vest in direct inference. (Kyburg's position is examined in detail in Chapter 7.)

The basic strategy in the reconstruction is to found fiducial inference on *significance testing*, just as in the N-P theory we find confidence intervals derived from hypothesis testing. Since the failures within the orthodox position stem from insensitivity to 'after-trial' analysis, it is crucial to the endeavor that significance tests avoid a similar fate. This is at least possible, since Fisher's solution to direct inference, i.e. selection of a reference class, incorporates an epistemological component in place of Neyman's pragmatic one.

In Section 3.2 the text offers a reading of Fisherian significance testing which *avoids* the repeated sampling interpretation advocated by Neyman (and examined in Chapter 2). The central considerations for the rebuilding are (a) determination of a *discrepancy ranking* for the test statistic (which summarizes the data), and (b) selection of a test statistic which includes *all and only* relevant features of the experimental outcomes. I define a simple discrepancy ranking D_p and consider its relation to a likelihood measure of support. Since likelihood support is overtly an 'after-trial' evaluation, this last concern is important for securing the desired properties of significance testing.

Using the ideas developed in Section 3.2, the three subsequent sections present criticism of several recommendations and objections raised in the literature surrounding significance testing. In 3.3 the discussion focuses on so-called 'data too good to be true,' and the counter-intuitive suggestions of such notables as Fisher and Jeffreys. In 3.4 I examine an objection to significance based on inconsistencies resulting from transformation of the test statistic used. This analysis is valuable for the defense of fiducial inference, upcoming in 5.5. What is at stake is a distinction between transformations which do and do not preserve the epistemic (informational) content of the test statistic actually obtained. Finally, in 3.5, I argue that Fisher commits a more subtle version of the error identified in 3.4. The problem involves satisfaction of constraint (b) (of 3.2), that is, using a statistic which reports all and only relevent features of the experimental data available. The issues developed in 3.4 and 3.5 reflect the concern for maintaining the relatively observational (as opposed to theoretical) character of the term serving as the test statistic.

The final section of this chapter attempts to generalize the reconstruction of 3.2 to include significance tests of composite null hypotheses. Particular attention is paid to Fisher's remarks with respect to tests of homogeneity for the 2 × 2 contingency table. The emphasis is on avoiding the repeated sampling interpretation of the significance test. It is shown that generally no such interpreatation is available when the test is of a composite hypothesis, or when the test employs an *ancillary* statistic (for conditioning)

to achieve satisfaction of (b). Both of these features are duplicated by fiducial probability in the corresponding situations. Hence, the way is paved for responding to those who object to fiducial inference because of its failure to live up to repeated sampling expectations. (see 4.3, 4.7, and 5.4.)

3.1. INTRODUCTION TO THE FISHERIAN THEORIES OF SIGNIFICANCE TESTING AND FIDUCIAL PROBABILITY

The preceding sections analyzing Neyman-Pearson theories of hypothesis testing and confidence intervals attempted to show that the underlying interpretation of probability as probability$_2$ permitted an integration of the two programs into a unified position. However, it was argued that the theory of confidence intervals proved to be inadequate for complete 'after trial' analysis; hence, severely diminishing the worth of the underlying hypothesis tests for 'after trial' analysis. Neither of these conclusions is new, and with regard to the second, the originators of the theories consciously rebuffed any such interpretation. When faced with the difficulties for 'after trial' evaluation, they emphasized the frequency based explication of probability, upheld be a repeated sampling technique for defining the FPS or reference class, and thereby argued for the imprioriety of the additional 'after trial' demands.

In the following chapters the Fisherian alternatives to the Neyman-Pearson program are advanced. It is argued that, by using additional inductive principles (besides 1–5 of pp. 7–10), true to the basic frequentist doctrines, a composite program (duplicating the N-P one) is available. Unlike the confidence interval concept, it is claimed that fiducial probability does stand up to 'after trial' analysis. However, this benefit is achieved at the cost of the simple 'repeated trial' interpretation.

Anyone who has read Fisher's last, and most controversial work, *Statistical Methods and Scientific Inference*, is well aware of the claim that the insight behind both the theory of significance testing (the Fisherian analogue of hypothesis testing) and fiducial probability (the substitute for confidence intervals) is a novel explication for probability, in accord with Frequentism. Fisher writes as though there is only one interpretation to be considered (his newly proposed one), although he does not state what appears to be the case; to wit, he adopts *two* interpretations and his suggestions deal with the confirmational or probability$_1$ concept alone.

That Fisher is somewhat confused about his own dissatisfaction with the limiting frequency interpretation is made evident by the objection he offers.

Indeed, I believe that a rather simple semantic confusion may be indicated as relevant

to the issues discussed, as soon as consideration is given to the meaning that the word probability must have to anyone so much practically interested as is a gambler, who, for example, stands to gain or lose money, in the event of an ace being thrown with a single die. To such a man the information supplied by a familiar mathematical statement such as: 'If a aces are thrown in n trials, the probability that the difference in absolute value between a/n and 1/6 shall exceed any positive value ϵ, however small, shall tend to zero as the number n is increased indefinitely', will seem not merely remote, but also incomplete and lacking in definiteness in its application to the particular throw in which he is interested. Indeed, by itself it says nothing about the throw.[1]

Now this retort to the limiting frequency school is not without foundation; after all, the direct inference rule is marred by the reference class problem. However, to base a rejection of the 'typical' interpretation on the observation that it does not apply explicitly to the single case is to overstate the value of the difficulty. Where Fisher is accurate is his intuition that the solution cannot be, as for Neyman-Pearson it is, a matter of either pragmatics or 'repeated sampling,' interpreted as routinely duplicating the experiment to form a sequence whose purported significance is to supply mathematical limits for the FPS. On the contrary, by including an *epistemological* condition in the selection of the reference class for direct inference he supplies the missing ingredient for avoiding the 'after trial' difficulties. Thus, whereas N-P explicated 'randomness' of the single trial by way of purely factual constraints, i.e. the repeated sampling property, Fisher adopts a decidedly subjective approach. He says,

> It is obvious, moreover, that many subsets of future throws, which may include his own, can be shown to give probabilities, in this sense, either greater or less than 1/6. Before the limiting ratio of the whole set can be accepted as applicable to a particular throw, a second condition must be satisfied, namely that before the die is cast no such subset can be *recognized*. This is a necessary and sufficient condition for the applicability of the limiting ratio of the entire aggregate of possible future throws as the probability of any one particular throw. On this condition we may think of a particular throw, or of a succession of throws, as a *random* sample from the aggregate, which is in this sense subjectively homogeneous and without recognizable stratification.[2]

The key parts to this vague interpretation of probability, separating it from the N-P one, are (i) the application to single events or single trials directly; and more importantly (ii) the use of an epistemological factor, through the 'recognizability' condition, to determine the appropriate reference class for direct inference. Thus, Fisher builds investigator's background knowledge into the meaning of his probability for single trials by requiring an ignorance condition as well as knowledge of a stochastic frequency or distribution. This is made clearer by the remarks that,

> This fundamental requirement for the applicability to individual cases of the concept of classical probability shows clearly the role both of well specified ignorance and of specific knowledge in a typical probability statement. It has been often recognized that any probability statement, being a rigorous statement involving uncertainty, has less factual content than a statement of complete ignorance. The *knowledge* required for such a statement refers to a well-defined aggregate, or population of possibilities within which the limiting frequency ratio must be exactly known. The necessary *ignorance* is specified by our inability to discriminate any of the different sub-aggregates having different frequency ratios, such as must always exist.[3]

Summarizing, the probability that an event e has property F is p_1 iff e is *known* to belong to a reference class of 'possible events' with proportion p_1 having property F, and such that *no subset* of this reference class having different proportions for F-ness *is known*, i.e. no *recognizable* subset is known. Under these conditions, e is said to be a random sample (or random member) from the hypothetical reference class with respect to the property F.

There are two obvious impediments to a satisfactory understanding of Fisher's scheme. First, the status and origins of the hypothetical reference classes remain obscure. Certainly they have little if any advantage over reference classes defined by the 'traditional' limiting frequency interpretation. Second, the 'recognizability' condition, intended to replace the N-P 'randomness' qualification, is left vague and without sufficient explanation to determine the scope of application. Does it include all direct inference cases? Does it include inverse inference as well?

Consider the reference class problem first. Without doubt, the hypothetical reference classes serve as the basis for an explication of the underlying probability$_2$ concept. However, as truth conditions they fare no better than von Mises' interpretation, or the traditional alternatives. To attribute any existential rank to a class of possible outcomes of a hypothetical experiment repeated *ad nauseam* to generate a limiting frequency, demands an ontological liberty in excess of plausibility. Even the N-P analysis, with all its empirical committments, begged off from a thorough explanation of 'randomness' (beyond the law of big numbers or 'practical certainty') when actual sequences of 'repeated trials' were not available. Therefore, it appears that Fisher's hypothetical reference classes will not do as a basis for a satisfactory explication of the probability$_2$ concept. However, this does not preclude their use as a model for probability$_2$, seen as a theoretical term.[4] Understood this way (which is a reasonable perspective, given the absence of a wholly adequate interpretation) Fisher's idea merits some consideration. In fact, using this weakened version of Fisher's idea, some of his remarks

in defense of *infinite* hypothetical reference classes become logical, in light of the 'recognizability' requirement.

For instance, he says that,

> The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability. If, in a Mendelian experiment, we say that the probability is one half that a mouse born of a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not in the hypothesis which we wish to consider; moreover, the probability may not always be a rational number. Being infinite the population is clearly hypothetical, for not only must the actual number produced by any parents be finite, but we might wish to consider the possibility that the probability should depend on the age of the parents, or their nutritional conditions. We can, however, imagine an unlimited number of mice produced upon the conditions of our experiment, that is, by similar parents, of the same age, in the same environment. The proportion of white mice in this imaginary population appears to be the actual meaning to be assigned to our statement of probability. Briefly, the hypothetical population is the conceptual resultant of the conditions which we are studying. The probability, like other statistical parameters, is a numerical characterization of that population.[5]

Thus, by using a *hypothetically infinite* reference class, an assumption of probabilistic independence between members of the reference class can be maintained even after seeing some of the other members. This independence is a usual feature of most statistical distributions, if not a regular presumption in many practical applications. For example, repeated trials on binomial chance set-ups are often supposed stochastically independent, as in repeated flips of a coin or rolls of a die. If the reference class were finite, then the outcomes for the first members would, if known, automatically alter the direct probability assignments to subsequent events (by the 'recognizability' condition). This probability$_1$ dependence among stochastically independent trials may be all right for inverse inference, e.g. when the limiting frequency in the hypothetical reference class is unknown, but it is not satisfactory for simple direct inference when independence is presupposed and the distribution assumed known.

Therefore, Fisher's use of infinite hypothetical reference classes is to be allowed, subject to the understanding that they serve a heuristic purpose; to wit, they aid in the examination of the important 'recognizability' condition. By no means are they to be infused with the added worth of offering truth conditions (and they definitely have little verificational value) for the probability$_2$ concept (stochastic probability). There is, however, a supplementary benefit to the use of the Fisherian hypothetical reference classes.

One reconstruction of the fiducial argument, due to H. Kyburg[6] (and reviewed in a later chapter), makes use of an 'epistemological' interpreatation of probability. For Kyburg, there is no extra stochastic ingredient to statistical information beyond *actual* frequencies. Like Fisher, Kyburg focuses on probability for single events, and also like Fisher the central issue is the determination of the reference class for direct inference. Therefore, bypassing the issue of the *infinitude* of the reference class (which Kyburg replaces with arbitrarily large, but finite ones), the Fisherian 'recognizability' condition is translated into a 'randomness' condition within Kyburg's system. It seems, then, that Kyburg's 'randomness' is designed to serve the very same purposes as Fisher's 'recognizability.'

This raises the second and, for this inquiry, most important question. What is the intended scope of the 'recognizability' constraint? Clearly, from Fisher's own writings, the direct inference problem, i.e. determination of the correct reference class, is to be resolved by it. Most surprisingly, according to Fisher, it also generates fiducial probability, which involves inverse inference, i.e. inference in the absence of a priori probability determined by the Frequency Principle.

As hinted at above, Kyburg's reconstruction of the fiducial argument duplicates Fisher's intentions, at least so far as to subsume both direct and inverse inference under one general regulation: 'recognizability' or 'randomness.' It is my belief, one argued for throughout the following discussion, that no single principle can satisfactorily accommodate all the demands required of an inductive logic while simultaneously permitting the broad range of arguments, including inverse inference, claimed to be valid by both Fisher and Kyburg. For example, Kyburg defends the violation of CC in his program. However, I argue that the failure of conditionalization extends to his reconstruction of fiducial inference, thereby nullifying it. Moreover, it appears that his 'randomness' conditions imply the failure of CC.[7]

Therefore, the reconstruction offered here, like Hacking's, augments the Frequency Principle and 'narrowest' reference class rule (seen as a specialization of TRE) with additional principles, directed especially at the fiducial argument. What remains of Fisher's hopes for a unified theory of both direct and inverse inference, as captured in fiducial probability, is the observation that the reconstructed argument uses the extra rules to convert inverse inference to direct inference by way of an irrelevance assumption: that the data available after the trial is run can be claimed as evidence irrelevant to particular direct inference conclusions, made before the data are available.

Before engaging in the reconstruction, it is to the point to review and develop some of Fisher's ideas on a different and more primitive version of 'inverse inference': significance testing. This is done for two reasons. It serves to reinforce the difference between the N-P theory of hypothesis testing and the Fisherian alternative, based on the disparate solutions of the reference class problem for direct inference. It also supplies the backdrop against which the additional principles, needed for the fiducial argument, are to be justified. In other words, the new rules used to reconstruct fiducial probability are defended (in part) by showing that they interlace the fiducial conclusions with those from significance theory (at least for one parameter problems under location parameter formulation).

So as not to mislead the reader into an expectation of a fully satisfactory resolution to the fiducial problem, let me say here that the reconstruction offered is adequate subject to a severe limitation (discussed on p. 133) to avoid a version of the paradox (discussed on p. 12) dealing with the Principle of Insufficient Reason. This restriction is not limited to the position taken here, but applies equally to Hacking's proposed reconstruction. Nonetheless, considerable attention is paid to developing the fiducial argument to cover multiparametric problems, as the great majority of published concern revolves about these contexts.

Despite limited progress towards achieving this goal, no discussion of Fisher's views on inverse inference would be nearly complete without including some comments on the problems of experimental design, notably, the question of *randomization* in design. This topic is covered in the final chapter, where it is approached from the somewhat novel perspective of inductive relevance. Besides linking up with Fisher's ideas on *estimation*, it allows for a consideration of *likelihood* as a relevance measure. Unfortunately, one of the conclusions reached is that the sort of inductive rules needed to salvage the fiducial argument, even under the limitations suggested (above), conflict with the use of randomization as offered by Fisher. Thus, a search for a fully general frequentist theory of inverse inference, in the Fisherian style, is held to be fruitless.

3.2. A RECONSTRUCTION OF FISHERIAN SIGNIFICANCE TESTS

In his classic text *Theory of Probability*, Sir Harold Jeffreys opens a chapter titled 'Frequency Definitions and Direct Methods' with an anecdote:

> Lord Mansfield gave the following advice to the newly-appointed Governor of a West

India island. 'There is no difficulty in deciding a case – only hear both sides patiently, then consider what you think justice requires, and decide accordingly; but never give reasons, for your judgment will probably be right, but your reasons will certainly be wrong."[8]

If one extends Lord Mansfield's advice to those who attempt statistical inference, then most of all, statisticians accustomed to using so-called simple tests of significance had best take heed. Considering the hodgepodge conglomeration of techniques and rules which make up the repertoire of standard statistical practice, no other procedure is at once both so natural and yet so susceptible to allegedly conclusive refutation.

The goal here is to sketch a reconstruction of Fisherian significance testing which preserves its intuitiveness, prepares its connection to fiducial probability, while also circumventing what appear to be three typical errors. Two of the mistakes are specific recommendations, from the highest authorities, on how to properly conduct and interpret the standard tests. The third is an argument which aims at an overthrow of the whole program on the grounds that contradictory conclusions are sanctioned.

The theme for the discussion is that, by using the hypothetico-deductive argument as a heuristic model, all three errors are seen to hinge on a faulty analysis of the term 'observation report.' To tip my hand in advance, the fallacies include (1) a misidentification of the observation report, (2) a misconstrual of what an observation report is, and (3) an innaccurate accounting of what the relevant aspects of the observational record are. The last two are further linked by being the result of failing to properly separate the hypothesis being tested from the secure background knowledge [BK]. Also, these last two have counterparts for the fiducial argument. Some preliminary grounds-clearing for fiducial inference results from the attention to these difficulties as they arise for significance testing.

To begin, fix the overall framework which motivates significance testing. Among frequentists, that is, among those who hold exclusively an objective interpretation of probability which is independent of the observer (probability$_2$) e.g. Neyman-Pearson, or among those modified dualists who allow for probability$_1$ statements but only when based on probability$_2$ evidence, e.g. Fisher, it is often the case that Bayes' theorem cannot be used to determine the probability for a hypothesis in light of data.

The problem stems from the absence of so-called 'prior' probabilities. For exmaple, if a newly bent coin has been flipped 30 times, with the resulting frequencies of 16 heads and 14 tails, Bayes' theorem cannot be used to determine a posterior probability for the hypothesis that the coin is nearly

fair unless an initial probability for the hypothesis is available. But the frequentist interpretation prohibits using a prior probability in this case since either (*a*) the genesis of the bent coin is unknown and non-stochastic, or (*b*) the background knowledge does not include enough information for the calculation of the prior probability, even if the bent coin has a stochastic origin; thus precluding justification of a specific prior probability assignment.

Not to be dismayed by this unfortunate turn of events, one might look for some alternative route to posterior probabilities, bypassing Bayes' theorem and its objectionable priors. Following this strategy, Frequentists of both sorts were led to (i) the theory of confidence intervals (Neyman-Pearson – discussed in the preceding chapter), and (ii) fiducial probability (Fisher – discussed in the subsequent chapters). Without reviewing the numerous controversies surrounding these concepts, let it suffice to say that they shoulder weighty objections. It would be wise to have some sort of back-up arguments. This is where statistical 'tests' come in. They play the role of a rudimentary form of inductive support which, both historically and epistemically back-up the more sophisticated arguments of confidence interval analysis (as demonstrated) and fiducial probability (as will be shown).

Now to the reconstruction. Suppose that a statistical hypothesis h_0 (the null hypothesis), up for consideration, postulates a frequency distribution for some observable, a random variable [*r.v.*], whose value is fixed by an experiment. That is, the null hypothesis provides a frequency distribution for the possible experimental outcomes, called the sample space [Ω].

The official Fisherian doctrine for the test (which is the interpretation to be followed here) goes like this:[9] Fix on a characterization of the members of Ω such that they can be ordered in increasing discrepancy with the null hypothesis. A minimum requirement of a discrepancy ordering D is that the probability, given h_0, of getting an extremely discrepant outcome is very low. The significance level of an outcome of the *r.v.*, x, is defined as: the probability under h_0 of getting an outcome at least as discrepant as the x observed. Thus, if the D ordering is well defined, a fixed significance level, denoted α (a probability value), divides the sample space into two subsets; call them C (the critical region) and $\sim C$ (the complement of C in Ω). The test proceeds by noting whether the actual experimental outcome falls in C, or not. If it does, then the null hypothesis is said to be rejected at the α significance level.

The Fisherian explanation for the test is that if the outcome is in C (at the α level), then either a 'rare' event has occurred or else h_0 is false.[10] The smaller the magnitude of α, the less the probability of getting an outcome in

the critical region and the 'rarer' the observed outcome. It is then argued that, given the disjunction, it is more plausible to accept the second disjunct, i.e. reject h_0; the rarity being an unsatisfactory accounting for the actual outcome.

The Fisherian doctrine continues by noting that, in contrast to the Neyman-Pearson program, (i) there is no mention made of alternatives to the null hypothesis, hence no 'power' associated with the test, and (ii) although there is a simple frequency interpretation which permits one to say that, on repeated trials, when h_0 is true it will be rejected falsely (type$_1$ error) in just $\alpha\%$ of the time, such a frequency property has little or no connection with the interpretation of the test.[11] To repeat, the correct interpretation is through the disjunction, either a rare event has occurred or the null hypothesis is false.

Criticisms of significance testing, so construed, abound. Several are serious enough to cast suspicion on the whole program. First, unless there is some interpersonal discrepancy ranking no significance test can be considered forceful, since there is bound to be some other ranking which, on the basis of the same experimental outcome and level, leaves the null hypothesis unrejected. To argue, as some do, that a different D ranking represents a different pragmatic concern of the investigator, confuses problems of practical decision with questions of inductive support. Where significance does play a part in decision, it is the level of rejection α which captures the pragmatic features.

In response to this criticism, which is a first cousin to the well known lottery paradox (as will be demonstrated), restrict the discrepancy measure in the following way.

DEFINITION: D_p (the probabilistic discrepancy ranking): Given h_0; the discrepancy of an outcome is inversely proportional to its probability (density, in the continuous cases). Hence, the higher the probability (density) of an outcome, given h_0, the lower its discrepancy. Those experimental outcomes with relatively the lowest probabilities (densities) are counted the most discrepant.

The justification for D_p is complex and cannot be given all at once. It is hoped that the reader will see the defense for this discrepancy ranking developing as the reconstruction for significance testing continues, especially whenever the tie between significance and fiducial probability is secured. The tie is limited to cases of location parameterization. A superficial summary defense is, however, available through the following important claim.

CLAIM: For location parameter problems the D_p ranking and the likelihood function agree. That is, with the random variable *fixed*, say at the value x_0, the relative likelihood for h_0 (against the composite alternative) coincides with the relative significance for x_0 (against different outcomes), using D_p and location parameterization.

Second criticism: even with a clear and objective notion of discrepancy D_p, if the sample space is altered either

(2*a*) by a 1-1 transformation of the sample space, or

(2*b*) by a modified description of the experimental outcomes (a reduction of the data), the resultant significance test for the same null hypothesis is *not* invariant. This two part objection will be the focus of attention for the rest of this section.

Consider first (2*b*). Suppose that, for example, the observational record is modified so that all results are rounded off after the third decimal place. Then, obviously, D_p applied to the new, reduced data fails to distinguish two outcomes which differed initially only beyond the fourth decimal place. If the unreduced data had unequal probability assignments, then before hand the two outcomes were of unequal discrepancy with the null hypothesis.

The problem cannot be satisfactorily answered by the simple retort that any reduction of the intiial data, as for example by rounding off, must be avoided. The most intuitive significance test of all relies on just such a maneuver. Suppose one tests the null hypothesis that a certain coin satisfies a binomial model with $\theta = .5$, i.e. a fair coin, when flipped by device F. The experiment consists of flipping the coin ten times and recording the sequence of heads and tails that results. Given h_0, each sequence of ten flips has exactly the same probability as any other. That is, the probability distribution is uniformly $\frac{1}{2}^{10}$ for each of the 2^{10} members of Ω. Hence, using D_p, all outcomes are equally discrepant with h_0, and significance is trivial. There are only two α levels; $\alpha = 0$, whence C (the critical region) is empty, and $\alpha = 1$, for which C is the whole sample space. Clearly, such a test has no value.

The usual test in this circumstance is based on a modified sample space generated by using as a random variable merely the proportion of heads (or tails). Thus, the new sample space is the eleven membered [0, 1] range; divided into intervals of $\frac{1}{10}$. The probability distribution for the modified *r.v.* is a uni-modal, symmetric 'bell-shaped' one with peak at .5. The resultant significance test is the familiar 'two-tailed' rejection region. The null hypothesis is rejected whenever the observed proportion differs by too much from the value .5.

The reduction of the data from the exact sequence to the frequency of heads represents contraction in information, just as that obtained from rounding-off. Hence, reductions cannot be ruled out, *prima facie*.

To help put the issue in focus, consider a heuristic analogue for significance tests: the elementary hypothetico-deductive argument. Imagine an investigator who is prepared to test a proposed hypothesis H which is independent of his background knowledge BK. For the experimental set-up E, BK conjoined with H yields the experimental consequence O (1*a*); or, equivalently (by the deduction theorem) (2*a*). (It is assumed, of course, that H appears essentially in the derivation of O.) If, in fact, the observation report is $\sim O$ (3*a*), the conclusion $\sim H$ (4*a*) quickly follows, and the investigator is, so to speak, justified in his rejection of H.

1*a*	$BK \,\&\, H \vdash O$	1*b*	$BK \,\&\, h_0 \underset{A}{\vdash} r.v. \in \sim C$
2*a*	$BK \vdash H \rightarrow O$	2*b*	$BK \underset{A}{\vdash} h_0$ "$\rightarrow$" $(r.v. \in \sim C)$
3*a*	$\sim O$	3*b*	$r.v. \in C$
4*a*	$\therefore \sim H$	4*b*	$\underset{A}{\therefore} \sim h_0$

The Heuristic Analogy

Care must be exercised when trying to cash in the analogy. First of all, because of the total relevant evidence requirement, for the statistical case it is important that the BK *not* include the 'observational report' that $r.v. \in C$. Then, the counterpart to 1*a* (which states that on the modified background $[BK \,\&\, H]$ O is a predicted outcome) must be that on the modified background $[BK \,\&\, h_0]$ it is predictable that the $r.v.$ will not fall in the critical region C. But $\sim C$ has a probability$_2$ (and by the direct inference, a probability$_1$) of only $1 - \alpha$, given $BK \,\&\, h_0$. Therefore, an extra acceptance step must be included to reach the desired prediction. Call the acceptance rule A, which is a 'high probability acceptance rule,' given that the magnitude of α is very low.

Next, there is no deduction theorem for conclusions arrived at through acceptance rules like A, so 2*b* must be reinterpreted. Specifically, the high probability for '$r.v. \in \sim C$' (from 1*b*) is now merely a *conditional* high probability, given BK and on the assumption of h_0. Hence, the prediction also goes conditional. But this requires a pun on 'conditional.' The material conditional of 2*a* is to be interpreted as the conditional of conditional probability.

Thirdly, the observation report $\sim O$ is replaced by the reported value of the *r.v.* which results from the experiment. By assumption, it falls in the critical region C. This leaves the investigator with three options only one of which leads to the desired conclusion; namely, the rejection of h_0. The invertigator can, possibly, deny the evidential status of the observation report. It might be that it occurred under abnormal conditions, e.g. that it is an hallucination. This alternative applies equally to the hypothetico-deductive argument and is dismissed, out of hand, for the 'normal' cases.

A second option, applicable just to the significance test, is the acceptance of the observation report without rejecting the null hypothesis. In other words, since the acceptance rule A is merely a 'high probability rule,' it could be admitted that a 'rare' event occurred, i.e. that h_0 is consistent with the observed experimental outcome. The formalization of this way out is by pointing to the TRE condition and arguing that, with knowledge of the experiment already available, BK must include the observation; hence, defeating the antecedent conditions for $1b$. Exactly when this excuse is plausible (from the Frequentist position) is hard to describe. An evenhanded, if not charitable evaluation is to see this option as a countervailing force against the claim that, in the 'normal' cases, rejection of the null hypothesis is mandatory. Above all, it is to be remembered that Fisherian significance testing is an extremely primitive form of inductive analysis.

Finally, when neither of the preceding two policies is adopted it becomes plausible to reject the null hypothesis. That is, retaining the conditional prediction of $2b$, in the face of $3b$, leads to the undesirable conclusion that, 'Were h_0 accepted then a contradiction would result.' Assuming that suspension of judgement is ruled out, there remains only the rejection of h_0. Hence, $4b$ follows, subject to use of the masked acceptance rule A and the various privisoes for alternative responses.

Two important notes:

1. Both $2a$ and $2b$ are essential since it is imperative that the hypothesis being tested be kept distinct from the secure background knowledge, BK.

2. A lottery-paradox charge stemming from the use of a high probability acceptance rule A can be avoided. Note that the rule actually employed is a function of the discrepancy measure D_p, which singles out a *unique* rejection region, if any at all.

Return, now, to reconsider the problem of the description of the experimental outcome which is to serve as the observation report, i.e. the *r.v.* In the deductive case the experimental result need be described only with respect to those features relevant to the hypothesis being tested, that is,

only with respect to properties common to the prediction O. So too for the statistical case. The *r.v.*, which is the counterpart for the observation report, must make use of only relevant features of the experimental outcome. In other words, the description of the experimental outcome need use *only* relevant properties, relevant to h_0.

Harking back to the fair coin example, it is demonstrable that, given a binomial assumption (in *BK*), the order of the outcomes within the 10-fold sequence of flips is irrelevant to a test for the bias. The *r.v.* used in this case, namely the proportion of heads observed, does not contain the extraneous, irrelevant information.

There is a breakdown in the analogy, however, when the converse question is considered. In the deductive case, if the observation report is compound, say conjunctive, as long as there is a conflict with the predicted outcome in at least one conjunct, the other features need not be considered. They remain relevant in the sense that they either confirm or conflict with corresponding conjuncts in the prediction. However, a contradiction with the prediction in even one conjunct is just as secure (logical) grounds for denying the hypothesis being tested as if a multiple conflict had resulted.

Not so in the statistical case where conflicts are only probabilistic. The ghost of the Total Relevant Evidence [TRE] requirement lingers on. The description, or *r.v.*, selected must report the outcome on all relevant fronts. That is, it is not enough to analyze the relevant properties singly, but in conjunction only. To sum up, the first relevance constraint demands that *only* relevant properties be used in the description. The second is, reminiscent of the TRE requirement, that *all* relevant properties be represented. Conjoining the two, the *r.v.* must be constructed from all and only relevant aspects of the experimental record. In more technical jargon, the *r.v.* must be the 'minimal-sufficient statistic' for the hypothesis tested.

Needless to say, some contraction of the original *r.v.*, i.e. the original description, is almost assured. Rarely is the experiment so efficient as not to produce extraneous information. On the other hand, reductions which throw away some of the relevant information are to be barred. Just how this condition is to be obeyed, and just what the role of *BK* is for determining relevance, will be answered when the third error is discussed (p. 90).

This completes enough of the reconstruction of significance testing to meet the needs in the discussion which follows. One should note that whereas the official doctrine leaves all of: (1) the discrepancy measure, (2) the *r.v.*, or observation report, and (3) level of significance α selected for rejection, free floating parameters, the reconstruction here fixes the discrepancy

ranking D_p in terms of the *r.v.* The determination of the random variable is, itself, a question of relevance; also a non-pragmatic issue. Also, the level of significance chosen for rejection is left open, as is traditional. However, the option remains of concluding the significance test by citing the level of significance of the actual experimental outcome, thus making external the problem of what the rejection level for a particular decision ought to be.

3.3. DATA TOO GOOD TO BE TRUE

The first mistake to be examined concerns a recommendation of how to interpret a significance test correctly. Although frequently discussed and nearly infamous for its results, this suggestion has a simple rebuttal given the reconstruction (above). The argument, which goes by the popular title 'Data too good to be true,' is given below.

An often encountered null hypothesis is the generalized multinomial distribution. As the name suggests, the multinomial distribution is related to the binomial distribution just as a multifaced die is related to a coin. If one represents each side of the hypothetical n-faced die by a cell, then a simple null hypothesis assigns to each of the n cells a frequency or expected value. For example, the null hypothesis that an ordinary die is fair corresponds to the assignments $\frac{1}{6}$ to each of the six cells, on a single roll.

The most widely used significance test for multinomial distributions is the χ^2 test for 'Goodness of Fit.' In the example χ^2 is defined as follows. Let the experiment consist of rolling the n-faced die $1000n$ times and recording the experimental frequencies for each cell. The cell_i component for χ^2 is:

$$(e_i - o_i)^2 / e_i;$$

where e_i is the expected frequency for cell_i, under h_0, and o_i is the observed frequency for the cell. Then, $\chi^2 = \Sigma_i (e_i - o_i)^2 / e_i$, with $n-1$ degrees of freedom. For example, if the null hypothesis is that the n-faced die is fair, then e_i equals 1000 for each cell_i.[12]

If the observed frequencies are far from the expected values, then χ^2 will be large and far off in the right hand tail, leading to rejection of the null hypothesis. But, notes Fisher (and in this he is backed up by Yule, Kendall and even Harold Jeffreys),[13]

> The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P [i.e. the smaller the value of χ^2], the more satisfactory is the hypothesis verified. . . . Occasionally small values of χ^2, beyond the expected range do occur. In these cases the hypothesis considered is as definitely disproved as if χ^2 had been very large.[14]

Perhaps the most famous application of this argument is by Fisher himself to Mendel's well known experiments on pea plants. In a 1936 article,[15] 'Has Mendel's Work Been Rediscovered?' Fisher observes that if Mendel's data are analyzed by Goodness of Fit, the observed frequencies conform so well to expected values that χ^2 is very low, well into the left tail of the χ^2 distribution, part of the rejection region. By the 'too good to be true' argument, the hypothesis tested, to wit: Mendel's well known 2:1, 3:1 and 9:3:3:1 laws must be considered refuted by his own experiments. Of course, Fisher does not propose that classical genetics be rejected, only that one should be prepared to accept an alternative explanation: that Mendel (or more probably some ambitious assistant) doctored the results to make them conform more strongly to the hypothetical predictions.[16]

Since, by the reconstruction given here, a two tailed test is the recommended one for 'bell-shaped' distributions, χ^2 included, at first blush it appears that Fisher *et al.* are correct. However, in this case intuitions are accurate and it is the experts who are mistaken. Their slip is a simple, but natural one. They have confused the description of the outcome, the *r.v.*, with the dsicrepancy measure. Goodness of Fit is one of the few cases where the D_p measure is *superficially* replaced by another one, in this case D_{χ^2}.

DEFINITION: D_{χ^2} (the χ^2 – Goodness of Fit discrepancy ranking): the higher the observed χ^2 the greater the discrepancy. The justification for the replacement is that D_{χ^2} is asymptotically consistent with the usual D_p ranking. In both cases the *r.v.* used is the multinomial cell frequencies, akin to the proportion of heads and tails in the intuitive binomial coin test. Using this *r.v.*, D_{χ^2} and D_p are mutually consistent. That this is so can be seen from the following simple argument.[17]

As the sample size increases (at least so large that no e_i remains less than say, .5) the multinomial distribution approaches the multivariate normal distribution. For this approximating multivariate normal the sum χ^2 can be represented as the sum of squares of $n-1$ independent random variables each normally distributed with 0 mean and unit variance, i.e. N(0, 1). The discrepancy ranking D_p applied to the multivariate normal approximating distribution can also be represented as the sum of $n-1$ independent, standardized normal random variables, each with a 'two-tailed' test, centered about 0. Therefore, as the observed χ_i^2 increases, D_p, applied to the corresponding cell factor, also increases; and conversely. (The restriction to minimum expected value is due to the *dependence* between the mean and variance for the multivariate distribution. As the $e_i \to 0$, so does the variance for cell$_i \to 0$.)

Practically, χ^2 is a readily tabulated distribution. By contrast, the full multinomial distribution would take up volumes and volumes to cover even the most elementary combinations. That Fisher also saw D_{χ^2} as a convenient substitute for D_p, while the latter was the correct discrepancy measure for exact, precise results, will be demonstrated in a later section, dealing with composite null hypotheses and the '2 x 2' contingency table.

How, then, can it be that the experts have mistaken the discrepancy measure for the observation report? What makes their mistake natural? The excuse is that χ^2 plays another important role in significance testing. Besides Goodness of Fit, χ^2 is used in the analysis of variance of the normal distribution, for instance. In this case the *r.v.* used has a χ^2 distribution (in fact, the *r.v.* is called the χ^2-statistic) and D_p is employed to yield a two-tailed test. That is, very low χ^2, for analysis of variance, is as discrepant with the null hypothesis as is very high values. Unfortunately, Fisher, in the classic text *Statistical Methods for Research Workers*, §21.02, lumps these problems under the same Goodness of Fit heading, further confusing the issue, but explaining (at least in his case) the source of the counter-intuitive 'too good to be true' recommendation.[18]

A final test for this explanation is to note the reaction of at least one of the experts to the special case, one degree of freedom. For this instance, and this one alone, χ^2 has an inverted *J*-shaped distribution. That is, it has a right hand tail distribution. If the explanation (given above) is correct, then the mistaken belief that χ^2, for Goodness of Fit, is the *r.v.* (not the *D* ordering) would lead to rejection *only* for high values of χ^2, unlike the rest of the 'too good to be true' analysis. Harold Jeffreys, one of the experts, says, "When there is only one degree of freedom to be tested, a very close agreement is not remarkable."[19] All of which is in support of the explanation given here!

The problem just discussed, although exciting in that it involves some of the foremost authorities on a subject of some controversy, is not central to the reconstruction of significance testing and fiducial inference. The source of the mistake is a peculiarity of the statisticians involved. Within the bounds of the reconstruction, their oversight is the confusion of the *r.v.*, or observable, with the discrepancy measure. There is not much philosophy in seeing that: good statistics, perhaps?

3.4. ILLEGITIMATE TRANSFORMATIONS OF THE RANDOM VARIABLE

The second problem to be considered has, in contrast with the first, important consequences for subsequent questions. Although this error has

been made several times in the ongoing story of significance testing, it seems to have been overlooked by the professionals. Reconsider objection (2*a*) (of p. 80), i.e.the use of 1-1 transformations to switch *r.v.*'s. It has been argued as recently as the September, 1974 volume of the *British Journal for the Philosophy of Science* that significance tests are inconsistent over such transformations.[20]

The reader is reminded that the foregoing (reconstructed) theory of significance testing, especially the appeal to the D_p, or 'probabilistic' discrepancy measure, is directed at location parameter formulations of null hypotheses. It is trivial that a transformation of a *r.v.* in location parameter form can produce one not in location parameter form. Since the motivation for the D_p measure (when location parameterization holds) is primarily the agreement with a likelihood measure of inductive support, a simple solution to the problem of transforming the *r.v.* is to modify the D ranking so as to preserve the agreement. However, the nature of the problem discussed in this section is considerably different. No objection is made to the criticism that a modified random variable requires a modified discrepancy ranking to maintain satisfactory results, i.e. agreement with a likelihood measure of support. But, taking this into account, it appears that certain procedures for generating transformations, for the purpose of switching random variables, are illegitimate. That is, notwithstanding the revision of the discrepancy measure, the two random variables are *not* equivalent 'observation reports.'

The authors Spielman and Redhead try to upset the theory of significance testing (in this instance, as formulated in D.A. Gillies' book, *An Objective Theory of Probability*)[21] by showing that 1-1 transformations of the random variable lead to a paradox: Regardless of the specific null hypothesis, using D_p, the original *r.v.* leads to a rejection of the null hypothesis iff the test using the transformed *r.v.* fails to reject. So, it is argued, one test rejects just in case the other does not, and all are based on the same initial data. What could be more unsatisfactory than 'a program which pervasively sanctions contradictory inferences?'[22]

The importance of the condition that the transformation be 1-1 is to bypass the reduction question (2*b*) (see p. 80). Intuitively, once the transformation is defined, it would seem that no information, relevant or otherwise, is lost by using either of the *r.v.*'s. Hence, the two tests ought to agree. The case in point, i.e. applying Gillies' rule, is somewhat complicated.

The example Gillies reviews and which motivates the Spielman-Redhead criticism is a paradox due to Neyman,[23] but which is an instance of problem

(2*b*), i.e. reduction of the data; not (2*a*)! Neither Spielman nor Redhead discern the separation between (2*a*) and (2*b*). In fact, the paradoxes they raise in rebuttal of Gillies' textual analysis (of Neyman's problem) are all instances of (2*a*), i.e. transformational problems. (The Neyman paradox, stemming from the choice of the *r.v.*, will be the subject of the third error: evaluating relevance.)

The defense from the Spielman-Redhead charge begins by considering an exaggerated form of their mistake. Reverting to the heuristic model (p. 81), suppose that one were to test the hypothesis *H*: All objects are monochromatic, specifically color *G*, when viewed under normal light. Given a somewhat weaker background than is commonly held, one can imagine a reasonable test based on the conditional prediction *O*: Any object viewed under normal light will appear *G*. So, if an object is observed to be not-*G*, once can reject *H*.

Imagine, now, an alternative test, where the observations undergo the following '1-1' transformation:

$$\mathbf{T}^*: o \rightarrow o^*,$$

defined as:

(1) if the object is normally seen to be *G*, put on a not-*G* filter and then 'look.'
(2) If the object is normally seen to be not-*G*, put on a *G* filter and then 'look.'

Assuming that observations o^* are still classified as 'viewed under normal light' (which, of course, is absurd), then *H* will be rejected by an *o* observation report just in case o^* leaves *H* unrejected. (As noted earlier, a modified critical region for the modified observation would permit consistency with the first test. But this is not the point of the objection, as seen below.) The obvious response is that $\mathbf{T}^*$ distorts observations by contaminating them with *H*, the hypothesis being tested. Hence, $\mathbf{T}^*$ is an illegitimate transformation and o^* is not equivalent to *o*.

By way of further illustration, suppose that a *r.v.*, *x*, follows the simple statistical (density) law that:

$$\mathrm{p}(x) = \begin{pmatrix} 1 & \text{for } \theta \leqslant x \leqslant \theta + 1 \\ 0 & \text{otherwise} \end{pmatrix}.$$

That is, *x* is uniformly distributed between θ and $\theta + 1$ (θ unknown). Let $\mathbf{T}'$

be the transformation which yields the new random variable x', defined by:

$$T': x - \theta = x'.$$

T' is a 1-1 transformation (assuming θ fixed) and is even differentiable. The *r.v.* x' has a known distribution, independent of θ. In fact, x' is uniformly distributed on the interval [0, 1]. However, x and x' are clearly not equivalent observables for testing hypotheses about θ. Knowing x is quite different from knowing x', *unless θ is already known.* That is, $x = 5$ is logically compatible with $\theta \in [4, 5]$ and not other θ values. But $x' = .5$, which is the transformed x value if $\theta = 4.5$, is logically compatible with any θ value.

To separate off such observation distorting transformations, draw the following distinction.

DEFINITION: Transformation T is said to be *epistemically calculable* iff T makes use of terms all of whose values are contained in (or implied by) the background knowledge *BK*. (These are also called *calculable on BK*, for short.) Otherwise, T is said to be incalculable. Example: $T': x \rightarrow x'$ (given above) is incalculable if θ is not already known.

The transformations used by Spielman and Redhead to generate their paradoxes all employ the cumulative distribution function [c.d.f.] $F(x)$, for the random variable x actually observed. To repeat, $F(x_0)$ is defined as the probability of getting an x value less than or equal to x_0. $F(x)$, or more properly $F(x, \theta)$, is a function of the unknown parameter whose value is merely postulated by the null hypothesis. In order that their transformations be 1-1, both Spielman and Redhead insert the parametric value from the h_0 tested. But this requires presupposing the hypothesis being tested for more than the legitimate purpose of drawing observational predictions. By their procedure, an incalculable transformation is employed for switching from one *r.v.* to another. It is to be remembered that calculability is with respect to *BK* and not *BK* & h_0.

To repeat, the null hypothesis has been allowed to contaminate the observations through the transformations. No doubt, the oversight stems from a mistaken view of the status of the null hypothesis. h_0 is not known or accepted as true; certainly, at least not prior to the test. The importance of line 2*b* in the reconstructed significance test (see p. 81) is, at last, made clear.

Therefore, one cannot make use of the value of the unknown parameter as posited in the null hypothesis to construct alternative observables, i.e. *r.v.*'s. Spielman and Redhead commit an error in their evaluation of what counts

as an observable. If one accepts the excuse that the fallacy grows out of a misunderstanding of the place of the null hypothesis in the test, then they should not be treated harshly. They only follow tradition on this point, as will be made clear in the next section, where criticism (2*b*) – reductions of the data – is discussed.[24]

3.5. MISTAKEN ANALYSES OF RELEVANCE AND THE R.V.

The third error to be discussed uses the same theme as the Spielman-Redhead mistake, only in a more subtle variation. Whereas theirs stems from an abuse of the null hypothesis leading to unsatisfactory 'observables,' the current problem arises by abusing the null hypothesis with resultant errors in the satisfaction of the relevance condition: that the *r.v.* use all and only relevant aspects of the original data. For illustration of this faulty analysis, look again at one of the foremost texts on significance testing: none other than Fisher's *Statistical Methods for Research Workers.*[25]

Up to this point in the reconstruction, the criteria by which relevance is determined has been bypassed. It has just been assumed that the *r.v.*'s used all and only relevant features of the experimental record. The insight needed to make sense of the error is the understanding that with respect to the null hypothesis, relevance is a question of the background knowledge, *BK*, also. A changed *BK* can lead to a reassessment of relevance for the same h_0.

Before considering an example from Fisher's own work, examine the following, intuitive case. To repeat the analysis of the coin test ot p. 80, it is claimed that if the *BK* includes the assumption of a binomial distribution (with unknown θ), a proper *r.v.* for a test of the bias, say that the coin is fair, does *not* include the order of the outcomes within the 10-fold sequence. However, if *BK* is modified so that instead of a general binomial assumption, the hypothesis that the coin is fair is just one of a set of disjuncts (all of which are logically consistent with *BK*) and one of which is the hypothesis that flips are strictly dependent, yielding the alternating sequence *HTHTHTHTHT*..., then eliminating the order from the observation report throws away some relevant evidence for testing the fairness against the other hypotheses. That is, using just the proportion of heads as evidence does not retain all the relevant information about h_0: that the coin is fair. Certainly, it is always the case that h_0 is tested against, at a minimum, $\sim h_0$. But, given *BK*, the alternatives are not merely the logical contradiction to h_0. Modifying *BK* changes the viable alternatives to h_0, and these must be considered when evaluating relevance.

The general problem of assessing inductive relevance is a very complex

one and definitely not a settled matter. As would be expected of a statistician of Fisher's eminence, he has a well developed theory of relevance. For a good summary of his ideas one would do well to read the famous 1922 paper 'on the Mathematical Foundations of Theoretical Statistics'[26] and the subsequent (condensed) 'Theory of Statistical Estimation.'[27] It is most unfortunate that these papers have been interpreted by non-Fisherians (most regularly by Neyman-Pearsonites) as a theory of Point Estimation qua *guestimates*. That is, a theory of how to construct good, precise point valued guesses about some unknown parameter. It is claimed here (and discussed in the last chapter, dealing also with problems of experimental design) that the correct interpretation for these works is to see them as part of a theory of relevance – a theory of how to reduce large bodies of data to concise summaries, without loss of *relevant* information.

If this interpretation is allowed, then the error is easy to describe. One high level criterion of Fisher's theory is the concept of *efficiency*. Without going into the technical meaning of the term now, its pertinence can be summarized by saying that if one *r.v.* has a higher efficiency rating than a second, with respect to some parameter, then the one with the higher efficiency is a better summary of the data (for hypotheses about the parameter). It is possible that the efficiency rating may vary with the value of the parameter. Table II (below) gives the efficiency rating for five *r.v.*'s with respect to possible values of a linkage parameter for two genes (in maize). *r.v.*'s T_3, T_4 and T_5 all have 100% efficiency ratings for all values of θ. The null hypothesis that the genes are independent corresponds to a proposed $\theta = 50$. For that value, T_2 is fully efficient, but for any other value its efficiency falls off, and it is a worthless summary if the genes are completely linked.

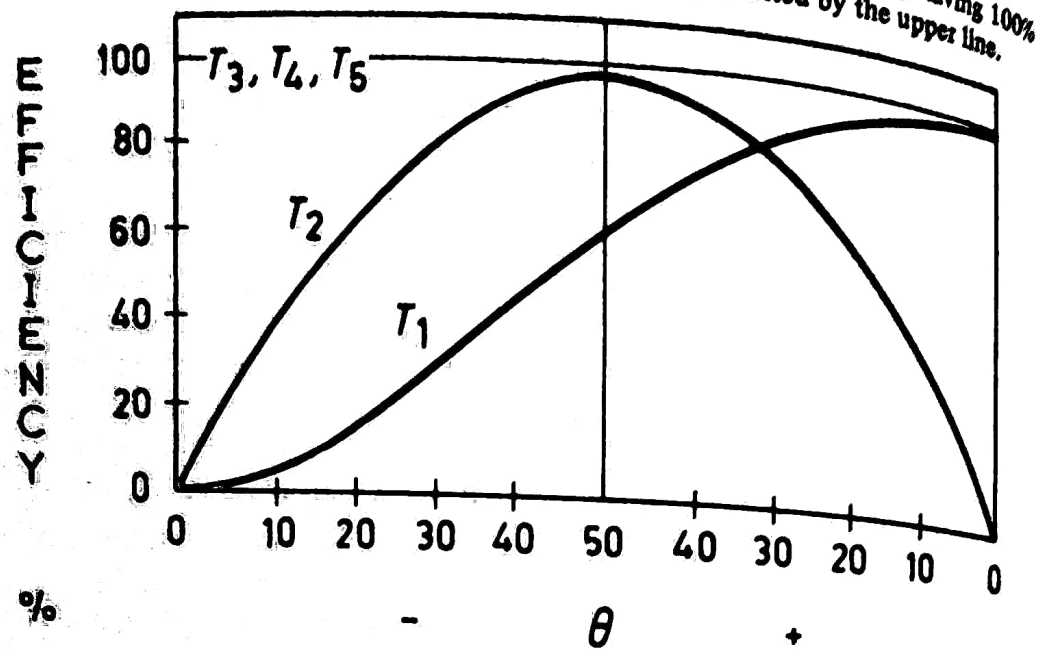
The third error is, as one can now guess, the recommendation that when testing the hypothesis that the genes are independently inherited, T_2 is just as good as a *r.v.* as one of T_3 through T_5. Fisher's own remarks are,

> T_2 is efficient when θ is $\frac{1}{4}$ [i.e. = 50 in Table II] or in the absence of linkage. This accords with the use of z [defined in terms of T_2] in Section 51 for testing the significance of linkage, for we are then testing the hypothesis that the factors are unlinked, . . . It is by no means uncommon to find statistics such as T_2 which provide excellent tests of significance, yet which become highly inefficient in estimating the magnitude of significant effect.[28]

What has happened is that Fisher has contaminated his analysis of relevance by assuming the null hypothesis. Were he accurate on this he would have realized that only T_3 through T_5 meet the relevance restriction; but T_2

TABLE II[29]

Efficiency of T_1 and T_2 for all values of θ. T_3, T_4 and T_5 having 100% efficiency throughout the range, are represented by the upper line.



does not, since it is only when the null hypothesis is true that T_2 is a proper summary. A modified version of the Spielman-Redhead paradox can be constructed using the faulty recommendation. To sketch it, consider two tests, one based on T_2 and the other on one of T_3 through T_5, and let h_0 be false but still compatible with *BK*, (say, the genes are strongly linked), then the conclusions can be mutually incompatible on the basis of the same initial data.

Fisher's own analysis demonstrates that when h_0 is true T_2 and T_3 through T_5 will be highly correlated ($\rho_{T_2\ T_{3,4,5}} \to 1$ as the sample size increases);[30] hence will be of equal 'observational'value. However, the relevance condition must take into account not merely *BK* & h_0, but *BK before* h_0 is accepted. As in the simple coin test example of p. 90, a modified *BK* leads to a different assessment of relevance.

Neyman's example, presented and partially discussed in Gillies' text, illustrates the problem of alternative reductions, i.e. competing *r.v.*'s. The motive Neyman had in constructing the 'paradox' was to demonstrate that significance tests could *not* be adequately formulated without appeal to specific alternatives to h_0. To permit a clear discussion of the issue let us review Neyman's example.

Suppose that n independent observations $x_1, \ldots, x_n$ of a normally distributed variable are made. *BK* contains only the information that the

variables are normally distributed and does not suggest values for either the unknown mean or variance. The null hypothesis to be tested states that the mean $\mu_x = 0$. This hypothesis, unlike those considered previously, is composite since it does not propose a value for the unknown variance σ_x^2. A discussion of composite null hypothesis testing follows this section, but in the meanwhile it can be said that the usual test for this null hypothesis is based on 'Student's' t-statistics, defined as:

Let $\bar{x} = 1/n \cdot \Sigma_i(x_i)$, the sample mean.
Let $s^2 = 1/(n-1) \cdot \Sigma_i(x_i - \bar{x})^2$, the sample variance.
Let $t = \sqrt{n} \cdot x/s$. Then, given h_0, t has Student's distribution, leading to a two-tailed rejection region (with $n-1$ *d.f.*). The conflicting test is defined by three different statistics:

Let $\bar{x}' = (x_1 - x_2)/\sqrt{2}$.
Let $s'^2 = 1/(n-1) \cdot \Sigma_i(x_i - \bar{x}')^2$
Let $t' = \sqrt{n} \cdot x'/s'$. Then, given h_0, t' has, just like t, a Student's distribution, leading to a two-tailed rejection region (with $n-1$ *d.f.*). However, it can be shown that:

$$|t \cdot t'| \leqslant 1.$$

From this it follows that the t-test rejects h_0 iff the t'-test does not, all on the basis of the same data $(x_1, \ldots, x_n)$.[31] Neyman resolves the question of which test is better through the corresponding 'power' analysis. For alternative hypothesis specifying non-zero means, i.e. against the alternative $\mu_x \neq 0$, the t-test is more powerful. Hence, Neyman argues, the alternative must be evoked to resolve the dilemma.

The Fisherian analysis of the same problem would go like this. First, both t and t' are calculable reductions of the data and therefore represent legitimate 'observables.' However, the two random variables do not fare equally with respect to 'relevance.' The two statistics $\bar{x}$ and s of the t-test are jointly sufficient for hypotheses about μ (in the absence of knowledge of σ^2). Not so for the competitors $\bar{x}'$ and s'. That is, the t'-test fails to make use of all the relevant information in $(x_1, \ldots, x_n)$ with respect to hypotheses about μ_x.

That the two analyses, i.e. 'power' and 'relevance,' agree in their recommendations is less surprising when it is remembered that UMP or UMPU-tests, when they exist, rely on sufficient statistics. That is, where power considerations pick out a 'best' test, as the t-test is (it is UMPU), the

rejection region is defined in terms of likelihood ratios, and those are constant along hyper-planes of constant value for the sufficient statistics. Therefore, the best tests are defined in terms of the sufficient statistics. Another perspective from which the agreement is less accidental is the sampling variance properties of the sufficient statistics. Where sufficient statistics exist they have a minimum sampling variance (among consistent normal estimators) leading to better 'power' by their use.[32] (All these results will be discussed in more detail in the final chapter where Fisherian theories of relevance are presented.) It is to be remembered, however, that these remarks presuppose a clear understanding of 'sufficiency,' and for multiparameter composite hypotheses such a presupposition is not fully warranted. (A more detailed discussion of this is found on pp. 137–40).

The upshot of all this analysis is that the Neyman-Pearson insistence on consideration of alternatives to the null hypothesis receives a round-about vindication when the problem of relevance is faced. One way of getting at the alternatives to h_0 is to review, prior to accepting h_0, the possibilities given just *BK*. This review, it has been argued, is necessary in order to fully conform to the 'all and only relevant properties' constraint (see p. 83). In other words, even for significance testing, relevance *is* a function of the alternatives to the null hypothesis, as set by the *BK*.[33]

It must be emphasized, however, that having granted the Neyman-Pearson point; namely, that proper tests cannot be constructed without considering alternatives to the hypothesis being tested, it does not follow that the two programs (hypothesis testing and significance testing) agree on the worthiness of tests, taken with respect to the relevance or reduction to a statistic factor. The hierarchy of criteria used to evaluate the operating characteristics of proposed tests do not automatically favor any test using a proper reduction, i.e. a minimal sufficient statistic, over one using less than a proper summary, i.e. an insufficent statistic. This follows because the N-P power considerations, through bias or power or invariance, include the evaluation of the discrepancy ranking of the corresponding significance test. Remember, though, that the official Fisherian doctrine leaves open the choice of discrepancy ranking, even after selecting a statistic. Therefore, the N-P warning to seriously examine the set of alternatives to h_0 is reasonable.

However, the N-P program rates unbiased tests and their correlated unbiased confidence intervals over biased ones. This can conflict with the recommendation of the relevance requirement if the discrepancy measure is ignored. For instance, suppose that the background information specifies that a sample of 9 readings are independently, normally distributed; each

with unit variance and unknown mean μ. The N-P UMPU-test for hypotheses about μ uses the sufficient statistic $\bar{x}$ (the sample average) and a 'two-tailed' critical region, defined by rejecting h_0 whenever $|\mu - \bar{x}| > \kappa$. However, compare the following two tests: (1) using the sample median m, and a 'two-tailed' rejection region, defined by rejecting h_0 whenever $|\mu - m| > \kappa'$ (which is unbiased), to (2) using the sample mean $\bar{x}$ and a 'one-tailed' rejection region, defined by rejecting h_0 whenever $(\mu - \bar{x}) \geqslant \kappa''$ (which is biased). Test (1) is preferred to (2), even though the median is a less than sufficient statistic for the unknown μ.

Both unbiased tests (the UMPU-test and (1)) use D_p (the probabilistic discrepancy ranking), however (1) violates the relevance condition. Nonetheless, Neyman-Pearson operating characteristics favor (1) (which is unbiased) over (2) (which is biased). If the difference in discrepancy rankings is ignored, the preference ordering for tests conflicts with relevance considerations. Of course, to repeat, the existence of a uniquely best test for both programs (the UMPU-test), results from the mutual agreement that D_p and Sufficiency, when conjoined, lead to optimum tests (for location parameter problems – otherwise D_p must be replaced by some other discrepancy measure with accords with D_L: the likelihood ranking). [See p. 87.]

3.6. COMPOSITE NULL HYPOTHESES, ANCILLARITY AND REPEATED SAMPLING

The last phase of the reconstruction of significance tests deals with the extension to composite null hypotheses. The principle conclusion argued for is that by appeal to the ancillarity irrelevance result (see p. 20), a complete break with the repeated sampling interpretation is achieved. This is exemplified for the so-called 2 x 2 contingency test.

Composite null hypotheses are characterized (typically) by the feature that the probability for the *r.v.* is not well defined, given the hypothesis. That is, just as in the N-P theory, the null hypothesis includes a set of possible parametric values, each member representing a different distribution over the sample space. The variety of composite null hypotheses is great indeed. For instance, the problem may by uni-parametric, with h_0 including a range of values, say $\theta \geqslant \kappa$. Or, the problem may contain a nuisance parameter, as in Neyman's problem discussed on p. 93. That is, the null hypothesis may specify a simple hypothesis for one parameter (μ) and leave the other (the variance) unrestricted. Or, the null hypothesis may restrict some of the parameters in a multivariate distribution. The combinations are endless.

The proposal made here is to extend the simple D_p discrepancy ranking,

to composite hypotheses, subject to similar restrictions. The intuitive motivation for this move is that a composite null hypothesis is interpreted as a 'disjunction' of simple hypotheses. Hence, the composite hypothesis is true so long as any of its 'disjuncts' are. Therefore, the composite discrepancy depends on the least discrepant component of the composite hypothesis.

DEFINITION: $D_{p\text{-max}}$ (discrepancy for composite hypotheses): The discrepancy of an observation x_0 with a composite null hypothesis h_0 is the *minimum* D_p value for x_0 over all the simple hypothesis making up h_0 (assuming location parameter form).

That is, the discrepancy for the null hypothesis is defined by considering one simple hypothesis (of h_0) such that x_0 has, subject to that simple hypothesis, the (relatively) highest probability (density). The connection between $D_{p\text{-max}}$ and likelihood is very clear for location parameter problems. In such cases $D_{p\text{-max}}$ is given simply by calculating probabilities (densities) using the maximum likelihood estimate (over h_0) of the parameter. In other words, by considering the simple hypothesis such that x_0 is assigned a maximum probability (density) from all those in h_0, the $D_{p\text{-max}}$ is obtained.

In the case of a simple null hypothesis, D_p induces an ordering of the sample space such that (with a continuous variable) the probability is exactly α (at the α-level) of obtaining an outcome in the critical region, given h_0. With composite null hypotheses, such is not the situation. A critical region consisting of all outcomes whose $D_{p\text{-max}}$ rating is α or less may fail to provide for outcomes such that all components of h_0 may be rejected at the α-level. That is, it may be impossible to obtain evidence which rejects all disjuncts of h_0 with equal cogency. In some cases, whenever h_0 is rejected at the α-level, specific disjuncts of h_0 can always be rejected at significance levels much lower than α.

On this point Fisher remarks that,

> Obvious as it might seem, it is evidently necessary to point out that it is no remedy to construct a test of significance with a firm intention that the hypothesis *shall* be rejected when true in a fixed proportion of trials. For this may well be mathematically impossible, for the whole range of cases; and consequently, what is of much greater importance, a test which is made to conform in one case may go widely astray in others. In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained should not be confused with the force, or cogency, of such evidence. It is the highest frequency, within the range of the class of hypotheses tested, that is logically relevant.[34]

An interesting, and important illustration of the difference between Fisherian and Neyman-Pearson significance testing, in the case of a composite null hypothesis, is the 2 x 2 contingency table. For this example, it is possible though the use of ancillarity to generate an exact significance test, i.e. one that rejects each element of the null hypothesis at the same level of significance, without supporting a Neyman-Pearson interpretation based on repeated samples from the same experimental set-up.

Suppose that two coins are flipped by a device F so that it is known that the flips are independent and binomial with unknown biases θ_1 (for coin_1) and θ_2 (for coin_2). The number of flips for each coin is determined by using a random number generator with integral values from 1 to 100. The composite null hypothesis is h_0: that $\theta_1 = \theta_2$.

Imagine the experiment run with the following results. The random number generator yielded 13 for coin_1 and 17 for coin_2. The outcomes of the 30 flips are recorded in Table 20.1.

TABLE 20.1[35]

	heads	tails	
coin_1	$10 = a$	$3 = b$	$13 = a + b$
coin_2	$2 = c$	$15 = d$	$17 = c + d$
	$12 = a + c$	$18 = b + d$	$30 = a + b + c + d$

From the conditions of the experiment, it is clear that the number of flips for each coin is determined by a process independent of the unknown biases. Therefore, (t(data) =) the right hand margin is a reduction of the data whose distribution is independent of the unknown parameters; hence is an ancillary statistic for the data.

If one concedes that the only relevant features of the experimental record are the two sets of heads/tails ratios, then the reduction of the data, t′(data) = the total number of heads/tails, i.e. the lower margin, is also ancillary (even given the value of t). Then both margins can be used as a single ancillary statistic. In other words, blotting out the four frequency counts, it is argued that the remaining data, i.e. the margins, supply no relevant evidence for testing the null hypothesis.

This maneuver has the pleasant consequence that, conditioning on the ancillary statistics, the probability distribution for the interior values, given h_0, is independent of the actual parametric value. That is, if $\theta_1 = \theta_2 = \theta$ and

the margins are held fixed, then the distribution for heads/tails within the table is independent of θ.

That this is so is evident from the following, elementary analysis:[36] the probability for the four outcomes, given the null hypothesis, is just

$$\frac{(a+b)!}{a!b!} \cdot \theta^a \cdot (1-\theta)^b \cdot \frac{(c+d)!}{c!d!} \cdot \theta^c \cdot (1-\theta)^d$$

$$= \frac{(a+b)!(c+d)!}{a!b!c!d!} \cdot \theta^{(a+c)} \cdot (1-\theta)^{(b+d)}.$$

But, with the lower margins fixed, the parametric factors are constant (no matter what θ), and holding the right margins constant keeps the probabilities for the four outcomes proportional to

$$1/a!b!c!d!.$$

Thus, by holding the margins constant, i.e. conditioning on the ancillary statistics, the $D_{p\text{-max}}$ ranking is independent of the unknown θ; hence, can be replaced by a simple D_p ranking.

This exact test, using D_p, can be duplicated by a Goodness of Fit test, using D_{χ^2} (see p. 85), with interesting sidelights. The χ^2-test requires an exact 'expected frequency,' given the null hypothesis. However, when the null hypothesis is composite, there is no precise expected value for each of the four frequency cells: e_a, e_b, e_c and e_d. This difficulty is circumvented be renewing the connection between $D_{p\text{-max}}$ and the maximum likelihood estimate of the data, given the range of h_0. (See p. 96.) That is, the $D_{p\text{-max}}$ is given by that value of the unknown parameter(s) (in this case, of θ) which maximizes the probability of the observed outcome (while conditioning on the two margins, i.e. the ancillary statistics). Therefore, the expected frequency based on h_0: $\theta_1 = \theta_2 = \theta$, is just the maximum likelihood estimate: which equals $(a+c)/n = 12/30 = .4$. That is, $\theta = .4$ is used to calculate the expected frequencies for the four cells a, b, c and d.

Another (historically) important problem connected with the χ^2 approximation for the 2×2 table is the question of the correct number of degrees of freedom for the test. If the exact test is kept in mind, so that conditioning on the marginal data is clear, then an entry for any of a, b, c or d automatically fixes the remaining three values; so that there is only 1 degree of freedom for the 2×2 χ^2-test of the composite hypothesis (with fixed margins).[37]

Finally, the failure of the exact test to support a simple repeated sampling

interpretation follows since, on subsequent trials, neither of the margins need be kept constant.[38] The distribution for the right margin is given by the specification of the randomizer; whereas that for the lower one is yet to be determined. In more usual 2 x 2 contingency tests the distribution of neither margin is clearly fixed by the experimental procedure. Therefore, for these cases, a repeated trial interpretation leaves both 'ancillary' statistics with unknown distributions. This is a very singular consequence. It is maintained that the two sets of marginal data provide no relevant information about the bias(es) of the two coins. Nonetheless, this irrelevance, based on the claim that the margins are ancillary for the bias, does not have a *precise* stochastic backing. In other words, it is *presumed* that the probability for the marginal data is independent of the unknown factor(s).

The somewhat special character of this assumption is emphasized by the fact that the presumption of *confirmational* irrelevance of the marginal data does not require a correlated presumption of *stochastic* irrelevance; although, of course, the additional premise is not ruled out. The sort of context where the marginal data are presumed irrelevant without a stochastic underpinning (by a direct inference: from stochastic irrelevance to confirmational irrelevance) is when there is no recognizable chance component in the determination of the margins. For example, modifying the coin example (above), allow that the number of flips for each coin might have been determined by some non-stochastic process, e.g. at the whim of the investigator. In spite of this, it might be argued that the marginal data provide no relevant information. Hence, the claim of ancillarity is made at the *confirmational level* without the support of a counterpart stochastic claim.

Apart from this, and even with stochastic ancillarity, conditioning on the ancillary statistic intervenes to destroy simple repeated sampling properties of the significance test. This foreshadows a parallel situation with fiducial probability. That is, just as Fisherian significance testing (which underpins fiducial probability) diverges from Neyman-Pearson hypothesis testing (which underpins confidence intervals) when ancillarity is used, so too the derivative fiducial probability and confidence interval theory diverges when the same principle is activated. (See p. 152.)

By way of illustration, all randomized tests from the N-P program are disqualified within the Fisherian theory, since the randomizer has (by assumption) stochastic independence from the unknown parameter(s). Hence, all data obtained from the randomizer are irrelevant, by the ancillarity rule, and cannot make a difference in the evaluation of inductive support. It follows that fiducial probability cannot appeal to randomization

to elimate ignorance, as Neyman does with his confidence interval solution to the Behrens-Fisher problem (see p. 63).

In conclusion, the theory of significance testing is not completed by the reconstruction given here. The continued restriction to location parameterization severely limits the application. However, the proximate goal was to lay the foundation for a tie between significance testing and fiducial theory; and for this purpose the use of location parameterizarion is no hinderance.

NOTES

[1] R.A. Fisher, *Statistical Methods and Scientific Inference*, 3rd ed. (New York: Hafner Press, 1973), p. 34.

[2] *Ibid.*, pp. 34–5.

[3] *Ibid.*

[4] As, for example, in R.B. Braithwaite, *Scientific Explanation* (New York: Harper & Brothers, 1953, rep. 1960), pp. 129–34.

[5] R.A. Fisher, Introduction to 'Theory of Statistical Estimation,' In *CMS*, p. 699a, prefatory note to paper #11.

[6] An early version of this reconstruction, emphasizing the 'epistemological' interpretation of probability is to be found in: H.E. Kyburg, Jr., 'Logical and Fiducial Probability,' Conference on Fiducial Probability, *Bull. Int. Stat. Inst.* 40[2] (1963), 884–901.

[7] This result is due to I. Levi, correspondence with H. Kyburg (December, 1972). I thank Professors Levi and Kyburg for having been shown the correspondence. Also, see I. Levi's 'Direct Inference,' *Journal of Philosophy*, LXXIV, #1 (1977), 5–29.

[8] Harold Jeffreys, *Theory of Probability*, 3rd ed. (Oxford: Clarendon Press, 1967), p. 369.

[9] This formulation of significance testing can be found in H. Cramer, *Mathematical Methods of Statistics* (Princeton: Princeton University Press, 1971), pp. 332–5. For a more recent statement see: F.J. Anscombe, 'Tests of Goodness of Fit', *Royal Stat. Soc., Ser. B* 25–6 (1963–4), 81–94.

[10] Fisher, *Statistical Methods and Scientific Inference*, p. 42.

[11] *Ibid.*, pp. 44–49.

[12] χ^2 (from Goodness of Fit) has a distribution closely approximated by the continuous χ^2 (from analysis of variance). That is, χ^2 has a dsitribution which is solely a function of the degrees of freedom [f]. For any moderate number of degrees of freedom, say $f > 30$, the distribution of χ^2 is closely approximated by observing that $\sqrt{2\chi^2}$ is Normal, with mean $\sqrt{2f-1}$ and unit variance. [This result is to be found in any basic test. For example, R.A. Fisher, *Statistical Methods for Research Workers*, 14th ed. (New York: Hafner Publishing, 1973), p. 61.] There is, of course, truncation at the left where χ^2 cannot be negative.

[13] G.U. Yule and M.G. Kendall, *An Introduction to the Theory of Statistics*, 12th ed. (London: 1940), p. 423; Jeffreys, *Theory of Probability*, pp. 307–10.

[14] Fisher, *Statistical Methods for Research Workers*, pp. 80–1.

[15] R.S. Fisher, 'Has Mendel's Work Been Rediscovered?', *Annals of Science* i (1936), 115–37.

[16] *Ibid.*, p. 132. It is to be noted that Fisher does not rest this conclusion solely on the weight of the 'too good to be true' argument. There is a more serious discrepancy with Mendel's results; see pp. 125–32.

[17] Again, this result appears in most texts; for example, Cramér, *Mathematical Methods of Statistics*, pp. 417–19.

[18] Fisher, *Statistical Methods for Research Workers*, p. 98.

[19] Jeffreys, *Theory of Probability*, p. 309.

[20] S. Spielman, 'On the Infirmities of Gillies's Rule', *Brit. J. Phil. Sci.* 25 (1974), 261–65; M.L.G. Redhead, 'On Neyman's Paradox and the Theory of Statistical Tests', *Brit. J. Phil. Sci.* 25 (1974), 265–71.

[21] D.A. Gillies, *An Objective Theory of Probability*, (London: Methuen, 1973), Chapter 11.

[22] Spielman, p. 261.

[23] Neyman's paradox is presented in Gillies, pp. 219–24.

[24] The attack on Gillies' rule, through transformations of the *r.v.* is not without foundation, although the criticism raised by Spielman and Redhead is not valid, as they present it. Gillies' rule uses, in effect, a constant discrepancy measure, D_p, for all statistical tests. No consideration is given to the parameterization of the null hypothesis. Therefore, by employing a *calculable* (non-linear) 1-1 transformation his rule can be shown to be inconsistent from the same initial data.

The lesson to be learned from this is, as noted previously, that no single discrepancy measure suffices for all tests. This does not renege on the earlier conclusion that, for a specific null hypothesis and background, only one D measure is proper. In other words, the discrepancy ranking is not arbitrarily or even pragmatically determined. Instead, the proposed reconstruction requires that the D ranking selected agree with a likelihood measure of inductive support.

[25] Fisher, *Statistical Methods for Research Workers*, pp. 300–22.

[26] R.A. Fisher, 'On the Mathematical Foundations of Theoretical Statistics,' in *CMS*, paper #10.

[27] R.A. Fisher, 'Theory of Statistical Estimation', *Proc. Camb. Phil. Soc.* 22 (1925), 700–725. Rep. in *CMS*, paper #11.

[28] Fisher, *Statistical Methods for Research Workers*, §56, p. 316.

[29] *Ibid.*, §55, p. 315.

[30] *Ibid.*, p. 316.

[31] The tests reject h_0 whenever the corresponding statistic (t or t') is not close to 0.

[32] For an early statement of this result see Fisher, 'Theory of Statistical Estimation,' pp. 714–20.

[33] For the N-P theory of hypothesis testing, it is not merely the alternatives to the null hypothesis which must be considered, but, also the aims of the inquiry, expressed by the way the alternatives are formulated, i.e. grouped togetner, affects the choice of test.

[34] Fisher, *Statistical Methods and Scientific Inference*, §5, pp. 95–96.

[35] Fisher, *Statistical Methods for Research Workers*, §21.01, p. 95.

[36] Fisher, *op. cit.*, p. 96. Also, Fisher, 'The Logic of Inductive Inference', *J. Royal Stat. Soc.* 98 (1935), 39–54. Rep. in *CMS*, paper #26; see pp. 48–51.

[37] Fisher, 'On the Interpretation of χ^2 from Contingency Tables and the Calculation of P', *J. Royal Stat. Soc.* 85 (1922), 87–94. Rep. in *CMS*, paper #5.

[38] A sequence of three articles by Barnard brings out some of the subtleties involved with the 'fixed margins' question: G. Barnard, 'Significance Tests for 2 × 2 Tables', *Biometrika* 34 (1946), 123–138; G. Barnard, 'The Meaning of a Significance Level', *Biometrika* 34 (1947), 179–182; G. Barnard, 'Statistical Inference', *J. Royal Stat. Soc., Ser. B* XI, #2 (1949), 115–140.

In Barnard (1946) the three standard models are presented: (i) both margins fixed; (ii) one margin fixed; and (iii) neither margin fixed. It would appear that a necessary condition for Fisher's assumption that the margins supply no 'relevant' information is that the margins be 'smooth' – without extreme values. The use of χ^2 presupposes no 'small' cell sizes, in any event.

CHAPTER 4

THE FIDUCIAL ARGUMENT: ONE PARAMETER

There are various possibilities for organizing a reconstruction of fiducial inference. One may, in perfect Fisherian style, present a cluster of numerical illustrations and then extract sufficiently many principles to cover them all without, it is hoped, also succumbing to contradictions. Alternatively, one may begin from a narrow base of well defined examples and work outward, establishing coherence and providing justification for the generalizations along the way. The text follows the second of these two options. Chapter 4 deals with one parameter problems. Chapter 5 extends the results of 4 to include multi-parameter fiducial inference, and concludes with a detailed categorization of, and resolutions to, outstanding 'paradoxes.'

Some explanation for this long (and sometimes technical) excursion into the reaches of fiducial inference is due the reader since, as becomes clear in 4.8, I believe the project of saving fiducial probability is doomed to failure. First, for those who have worried about making sense of Fisher's thirty-two years' grappling with the concept, it is important to understand where his intuition goes amiss. Surprisingly, I believe, it is not the weight of numerous technical difficulties ('artificial' would, no doubt, have been Fisher's description) which brings down the program. Instead, it appears that a fundamental problem, namely, giving a coherent representation to our *ingorance*, cannot be overcome. The point is made quite simply (4.8). Second, putting this objection aside, what is the status of existing criticism of Fisher's proposals? That is, what is the connection between the objection raised in 4.8 and the numerous 'counter-examples' already in the literature? My answer is that most can be rebutted; and where difficulties remain, they stand as a challenge to any program of inference which hopes to address the same questions.

The reconstruction begins (4.2) with a discussion of the role of the *pivotal variable*. Following I. Hacking, I argue that a fiducial inference obtains from a direct inference with the pivotal, *before the experimental outcome is known*, and then retaining these probabilities after the datum is available. What is required, clearly, is a set of restrictions on those pivotal variables which permit this last step, involving an irrelevance assumption. That is, we must identify the kinds of pivotal relations which allow one to say 'after-the-trial' that the datum is irrelevant to the 'before-trial' direct inference.

As a first move, I adopt J. Tukey's *smooth invertibility* condition on pivotals. (See Appendix 9.2 for a discussion of the equivalence of Tukey's constraint to one advocated by Fisher.) An example shows this to be a necessary condition for defending the irrelevance step in the fiducial argument.

Until this point in the analysis, it has been assumed that the experimental outcome(s) admit summarization in a single sufficient statistic. This overly restrictive premise, though of heuristic value for the opening discussion, must be eliminated if fiducial inference is to be of more than purely theoretical interest; too few actual problems satisfy the assumption to make it worthwhile pursuing such a limited theory. Therefore, I examine a widening of the smooth invertibility condition to allow for *exhaustive*

estimation (conditioning on an ancillary statistic) (4.3). The justification for the claim that this liberalizing move risks no new dangers makes use of confirmational conditionalization (condition (ii) of 1.2) in such a way as to suggest how fiducial inference may be applied quite generally. Specifically, by noting that conditionalization entails that inference must be invariant over all different sequences of acceptance leading to the same body of total evidence, one can provide a complete analysis of fiducial inference by examining three kinds of evidence pairs.

That is, one can begin with data for which a fiducial argument can be constructed (subject to the revised pivotal conditions of 4.3) and show how to incorporate new data which, on their own, fail to support a fiducial inference, see Section 4.4. Or, the supplementary data may, by themselves, also satisfy pivotal conditions for fiducial probability, and, moreover, the combined data may be of a kind to generate fiducial probability; see Section 4.5. Finally, the subsequent data may support fiducial inference but the combined data may fail to satisfy pivotal conditions; see Section 4.8. When we evaluate fiducial inference with respect to each of these cases, we obtain a complete account of the soundness of the argument for problems with exactly one unknown factor.

Using just the liberalized pivotal constraints developed in 4.3, I reconstruct Fisher's argument for fiducial inference with data 'of two kinds,' (4.4). However, for the second case, i.e. with data such that each part alone can support fiducial inference and a combined sufficient statistic also exists on which a pivotal can be built, D.V. Lindley's important analysis reveals serious drawbacks in the proposed inference scheme. Specifically, because of the nature of the evidence (and especially because the data comprise a pair) *three* different fiducial inference patterns can be constructed. That is, either part of the data may be used first in order to arrive at a fiducial probability and then the residual information added on by use of conditionalization (this gives two possibilities), or the total evidence may be summarized by a sufficient statistic and the argument developed without any sequential methods at all. Again, symmetry in Bayes' theorem demands that all three methods result in the same posterior probability, given all the data. Lindley constructs an example where each method yields a distinct answer, (4.5–4.6).

Fisher's polemical response to Lindley's objection is considered and rejected for lack of significant content, 4.5. Instead, in 4.6, I introduce an additional constraint on pivotal variables, i.e. the *canonical* pivotal, and argue that it gives a satisfactory answer to Lindley's problem. In fact, I examine three aspects of the modified argument (using smoothly-invertible, canonical pivotals) for justifying the emendation. First, by appeal to Lindley's necessary and sufficient conditions for coherent fiducial inference when the three inference patterns are available, it follows that no inconsistency can result in this case. Second, the modification does not appeal to unwarranted 'prior' probability. Finally, in light of an earlier (3.1) suggestion to link fiducial probability and significance testing, and by using the added canonicity limitation, we may interpret fiducial probability intervals as sets of unrejected null hypotheses (where rejection is by a Fisherian significance test). Two remarks which clarify this connection are presented in 4.7. In particular, the tie to *likelihood ratio* intervals is shown by way of the D_p (Section 3.2) discrepancy ranking.

Whatever hopes there are for a satisfactory reconstruction of fiducial inference, that have grown from the successes in preceding sections, are dashed by developments in Section 4.8. The last of the three cases (one not covered by Lindley's results), reveals

the inherent weakness of Fisher's project. A rather simple counter-example points out that the fiducial argument cannot provide a unique representation for 'ignorance,' that is, the argument suffers from the same infirmity as does the basic Laplacean rule of Insufficient Reason. To wit: a suitable reformulation (non-linear transformation) of the space of possible states, leads to mutually inconsistent solutions. Short of saving the argument by putting it in a Procrustean bed of artificial (case-by-case) limitations – which would cut off any hopes for applying the theory – no known remedy exists. As shown in 6.1, even Hacking's version of fiducial inference succumbs to the same weakness.

4.1. INTRODUCTION TO THE FIDUCIAL ARGUMENT

In 1930 R.A. Fisher presented the fiducial argument as an alternative method for obtaining statements of inverse probability in the absence of acknowledged prior probabilities.[1] As has been shown previously, frequentists of the tradition of Boole, Venn and Chrystal objected to the use of Bayes' theorem for this purpose since the prior probability factor had no frequency basis. In the preceding section, it was argued that significance tests took the place of inverse probability arguments. Their advantage was that the need for prior probabilities was circumvented, while retaining a clear frequency based interpretation. Their faults were also obvious. Primary amongst these was the inability to supply the desired conclusion, a probability for the unknown parameter in the light of experimental data.

Before presenting the reconstructed fiducial argument, with all its many details, it is best to review the historic 1930 paper so as to strengthen the tie to significance. Let us suppose that the background knowledge *BK* puts little or no constraints on the possible values of the unknown parameter θ. By that is meant that the background knowledge is insufficient for the determination of a frequency based prior probability for θ. Furthermore, under experimental circumstances *E*, *BK* is adequate for the conditional probabilities $\mathrm{p}(x/\theta)$. That is, under the assumption of a value for θ, frequency based probabilities for the sample space are defined. If x, the random variable for the experiment, is continuous the (confirmational) probability density function $\mathrm{p}(x/\theta)$ is obtained by the Frequency Principle from the c.d.f. $\mathrm{F}(x, \theta)$ in the usual way. If p is reasonably smooth, say 'one tailed' or 'two tailed,' then the probabilistic discrepancy ranking D_p (see p. 79) coincides with a simple function of F. For example, in the 'one-tailed' case D_p is either identical to the ordering (i) all values less than the observed outcome, or (ii) all values greater than the observed outcome. But (i) is the set used to determine F, and (ii) is just $1 - \mathrm{F}$. Thus, for the one-tailed distribution, significance and probability are numerically the same (or else sum to unity).

Considering 'two-tailed' distributions, once the additional assumption of symmetry is imposed, D_p and F are again simply related. In this instance, discrepancy for a given outcome is just a pair of 'tail' regions, each of equal probability. That is, the D_p picks out a subset of Ω equal to twice F, which in this case is defined by using only one of the 'tails.'

Returning to Fisher's paper, if $F(x, \theta)$ is interpreted as a function of two variables, not just a function of x given θ, then values of F can be treated as a *random variable* x_F with a known distribution. Values of F are uniformly distributed on the unit interval [0, 1]. So, with probability .95 $0 \leqslant x_F \leqslant .95$. Maintaining this probability statement about x_F even after x is observed induces a fiducial probability for θ given x. In terms of significance testing, if each possible value of θ is treated as a separate null hypothesis, then the set of θ not rejected at a significance level α (given an observed x) have a fiducial probability of $1 - \alpha$ (given the same data). In fact, his 1930 presentation gives tables for fiducial probability in terms of fiducial 'intervals.' These intervals are the surviving values of θ left unrejected by the corresponding significance test.

Fisher's claim of 1930, namely that the fiducial probabilities supplanted inverse inference based on unwarranted prior probabilities, was not widely accepted. The fiducial argument was no argument, but merely a poorly sketched technique. Even the tie to significance testing, a somewhat more popular method, was left vague. Because Fisher alleged a frequency based probability, his position was easily misrepresented. Granted the intimacy between fiducial inference and significance testing, and that by 1930 significance was *not* (for Fisher) interpreted by the appeal to repeated trials of the same experimental set-up, the ensuing confusion generated by linking fiducial probability with confidence intervals (which were so interpreted) was enormous.[2]

The problems with fiducial probability, both then and now, fall under several headings. What, exactly, is the fiducial argument? What are the necessary and sufficient conditions for its derivation? What is its frequentist basis? How, if at all, does fiducial probability differ from ordinary inverse inference, i.e. inverse inference using approved, frequency based prior probabilities? How does fiducial probability mesh with other, non-fiducial probabilities? Thus, the concerns involve not merely explication of fiducial probability, but also justification and scope of application within the probability calculus.

In responding to these demands, two positions will be considered. First, Fisher's response(s) (if he changed his mind radically) will be presented.

Second, an alternative to the Fisherian interpretation will be proposed. For the most part, these alternative suggestions represent improvements (many consistent with Fisher's own analysis). The need for the improvements will be made clear, in part by the discussion of several paradoxes raised against the fiducial argument, in part by the lapses in Fisher's reconstruction and explanation.

The reader is reminded that, in my opinion the fiducial argument is *not* salvageable, even for simple one parameter distributions. However, there is much to be gleaned from the analysis. Appealing to some remarks of L. Savage, "one possible definition of the fiducial argument is *'learning from Fisher.'* So interpreted, the fiducial argument is certainly valuable."[3]

4.2. THE PIVOTAL VARIABLE AND THE SMOOTH INVERTIBILITY CONDITION

To begin, consider the simplest case of a single (one dimensional) unknown parameter θ. As before, the background knowledge *BK* is too weak for the determination of a 'prior' probability for θ with the use of the Frequency Principle directly. There exists an experimental situation such that the conditional probabilities, given θ, for the sample space are defined. It is assumed that the sample space can be 'reduced' to a simple (one-dimensional) random variable without loss of relevant information about θ. That is, a sufficient statistic for θ is admitted. To sidestep this premise (for the moment) assume that the experiment is single-trialed, so that the observed random variable is already both sufficient and the proper dimension for θ.

The crux of the reconstruction is the selection of a hypothetical random variable, called a 'pivotal variable,' which is the focal point of the argument. The purpose of the pivotal is to allow, before the data are observed, certain frequency based probabilities which if retained after the experiment is run and the outcome observed, induce fiducial probabilities for the unknown parameter. All this proceeds, supposedly, without any hidden assumptions, to wit, an unwarranted use of a 'prior' probability. To further illustrate the argument and to emphasize the Fisherian constraints on the pivotal, examine the following two instances. One, the continuous example, is true to Fisher's admonition that the fiducial argument applies only to continuous random variables and parameters. The second, discrete binomial case is due to Ian Hacking, and although not in accord with Fisher's warning, is, to use Hacking's phrase, 'probably too insignificant' to have been considered by Fisher.[4]

Continuous

BK: Random variable x is N(μ, 1) with μ unknown, i.e. normally distributed about μ, variance = 1.

Pivotal: Define the hypothetical random variable w: $w = x - \theta$.

Pivotal probability: w is N(0, 1) independently of the unknown θ. (See Appendix 9.1.)

Fiducial step: If w remains confirmationally N(0, 1) after x is observed then θ is fiducially N(x, 1). Thus $p(\theta/x)$, the fiducial density for θ, equals $-\partial F/\partial\theta$.

Discrete

BK: A binomial coin ($\theta = .4$ or $\theta = .6$) is flipped once with random variable 'heads' or 'tails.' $p(\text{'heads'}/\theta = .4) = .4$

Pivotal: Define the hypothetical random variable o:

$o = o_1$ iff ($\theta = .6$ & heads) or ($\theta = .4$ & tails)

$o = o_2$ iff ($\theta = .4$ & heads) or ($\theta = .6$ & tails).

Pivotal probability: o is binomial $p(o = o_1/BK) = .6$, $p(o = o_2/BK) = .4$, independently of the unknown θ.

Fiducial step: If confirmationally o is binomial after the coin is flipped, then θ is fiducially binomial with

$$p(\theta = .6/\text{heads} \,\&\, BK) = p(\theta = .4/\text{tails} \,\&\, BK) = .6$$

and $$p(\theta = .6/\text{tails} \,\&\, BK) = p(\theta = .4/\text{heads} \,\&\, BK) = .4.$$

The crucial part of the argument is, of course, granting the antecedent of the Fiducial step, so that the fiducial probability can be detached. In order to make the final assumption benign, Fisher pointed out that the argument has the form of a *direct* probability statement, based on the pivotal probability step. In other words, since the pivotal probability statement is a direct one, by keeping the same reference class even after the random variable is observed, the antecedent for the Fiducial step is forthcoming. What remains to be shown is that no new relevance principle, besides 'recognizability,' is needed to retain the original reference class for solving direct inference about the pivotal *after* the data are observed.

Continuing in this vein, reconsider the Fisherian interpretation of probability. How, for example, is it that the probability of the next flip of this fair coin landing 'heads' is .5? The next flip is a member of the hypothetically infinite class of flips of this coin, which has (from the 'fairness' statement) a stochastic measure of $\frac{1}{2}$ 'heads' and $\frac{1}{2}$ 'tails.' The background knowledge is such that no *relevant* subset of this reference class is known to contain the next flip, so that it is a *random* member of the hypothetically infinite reference class with respect to landing 'heads up.' Thus it is given a .5 probability of landing 'heads up.'

What makes this infinite reference class the appropriate one? Why is no subset a relevant subset? It is in spelling out this answer that Fisher finds

the *already existing* principles which serve to restrict the possible pivotal variables for fiducial inference. The covering rule is that the reference class selected must be narrowest recognizable one, where 'recognizability' has the following properties:[5]

i – If a subset S of a given reference class R (known to contain entity e) has different stochastic measures than R for the limiting proportion of F's but it is *not* known whether e is in S, then S is *not* a recognizable reference class for inference about e's status as F or as not-F.

ii – If, with e and R as above, S' is a known subset of R and it is known that e is in S' but the stochastic measures for the limiting proportion of F's is the same in both sets, then S' is not a recognizable reference class for inference about e's status as F or as not-F.

Two examples illustrate these intuitive requirements. Suppose a potentially relevant reference class is determined as follows. Let S be the hypothetically infinite set of flips with this fair coin which land with the same side showing as the next flip. S is either the set of flips with this fair coin which all land 'heads up' or the set of flips which all land 'tails up.' Either possibility has frequencies known to differ from the .5 stochastic measure, in the hypothetically infinite reference class R of all flips with this fair coin, for landing 'heads up.' But no matter which alternative is selected for S, it is assumed that it is *not* known that the next flip is one of its members. Therefore S is not a recognizable reference class for inference about the next flip.

Second, suppose that the potentially relevant reference class S' is determined as: the hypothetically infinite set of flips performed on the same day of the week as the next flip. S' has stochastic measures which are assumed to agree with those for R (R as above). The next flip is known to belong to S', by definition of S'; however, S' is not a recognizable reference class for inference about the next flip.

How does recognizability affect the selection of a pivotal variable? It is through the requirement that, after the random variable is observed, the surviving reference class be either (*a*) identical to the original reference class for the pivotal probability statements, or (*b*) such that the stochastic measures agree exactly with the original reference class if (*a*) fails.[6]

To achieve this end, Fisher imposed two conditions on the pivotal variable. First, to guarantee the existence of the pivotal probability statements (by the direct inference), the pivotal variable (a function of the observed random variable and unknown parameter) has to have a calculable distribution which is independent of the unknown parameter, i.e. a known distribution (see

Appendix 9.1). To motivate the second restriction, whose purpose is to prevent recognizability after the random variable is observed, consider how the first condition alone is insufficient.

One hypothetical random variable, whose distribution is independent of the unknown parameter and which meets the other provisions of the first condition, is x_F the value of the distribution function, $F(x, \theta)$, itself. $F(x_0, \theta)$ is the probability, given θ, of obtaining an outcome smaller than x_0. For any value of θ, x_F is uniformly distributed on the interval [0, 1]. Since this distribution is independent of θ, it does not matter if θ is not constant from trial to trial. More precisely, the distribution for x_F is uniform on [0, 1], no matter how θ is distributed, i.e. no matter what the 'prior' for θ. It remains to show that probability$_1$ statements about x_F (obtained by a direct inference), do not always use the same (or equivalent) reference class both before and after observing x.

Suppose $F(x, \theta)$ is defined by the simple distribution law:

$$F(x, \theta) = \begin{bmatrix} x/\theta \text{ for } 0 \leqslant x \leqslant \theta; \quad \theta > 0. \\ 0 \text{ otherwise.}^7 \end{bmatrix}$$

That is, x is uniformly distributed between 0 and θ, θ positive; with density function $\partial F/\partial x = 1/\theta$ (for the positive sample space). x_F is, itself, uniformly distributed on [0, 1], so that, prior to seeing x the probability that $0 \leqslant x_F \leqslant .5$ is just $\frac{1}{2}$. Letting $x = 10$, and assuming the Fiducial step (see p. 108), the fiducial probability is $\frac{1}{2}$ that $0 \leqslant x/\theta \leqslant .5$, or equivalently, $\theta \geqslant 2x$: which on substituting the observed x yields, $\theta \geqslant 20$. Now imagine a change where the background knowledge is modified so that an upper bound for θ is known, say $\theta \leqslant 15$. Still x_F is pivotal, and is uniformly distributed on [0, 1]. However, since $\theta \leqslant 15$ it is incorrect that, after seeing $x = 10$, the probability is $\frac{1}{2}$ that $\theta \geqslant 20$, in fact, it is zero. What has gone wrong is that the new reference class for probability statements about x_F (in light of the observed x) takes into account the upper bound, unlike the initial reference class based on an unseen x. Note also that the modified background knowledge remains insufficient for determining prior probabilities for θ. Thus, Fisher's 1st concern: the existence of the pivotal, fails to distinguish the several cases where, in light of the experimental outcome, the reference class for probability statements about x_F changes to preclude the fiducial step; all in the absence of 'prior' probabilities.

Fisher reacted to the deficiencies in the pivotal by demanding that only 'natural' pivotals be used. As could be guessed, naturalness was just a blanket

to cover the many ways 'recognizability' might be avoided. Eventually, a more precise and informative restriction emerged. It was required that the pivotal, assumed continuous and differentiable in both parameter and variable, be (i) unbounded for all possible values of the random variable, and (ii) monotonic in both the parameter and variable.[8] It was supposed that these constraints only clarified the 'recognizability' assumption, without adding any fundamentally new principle; certainly not a hidden 'prior probability' assumption.

To illustrate the added constraint, review the continuous fiducial argument of a few pages back. The pivotal variable w, defined as $w = x - \theta$, satisfies both (i) and (ii). Being continuous and differentiable in both x and θ, it is to be noted that (i) with x fixed, w is unbounded, since θ is. (ii) If x is fixed, then w is increasing for decreasing θ, and if θ is fixed, then w is increasing for increasing x; hence, the pivotal is monotonic in both variable and parameter.

Before analyzing the discrete example to ascertain whether it stands up to the several pivotal requirements, it will be profitable to reformulate the 'recognizability' condition. Using J.W. Tukey's definition, a pivotal is said to be "smoothly invertible" if,

> "when any possible set of observations is inserted as arguments, the mapping from parameters to (pivotal) quantities:
>
> (*A*) has the same range for any possible set of observations,
> (*B*) is 1 to 1, and hence has a single valued inverse, and
> (*C*) this inverse is continuous (and has continuous derivatives)."[9]

It is easy enough to verify that the discrete example of the fiducial argument satisfies the two conditions. 1-existence of the pivotal, and 2-'smooth invertibility,' the 'recognizability constraint.' From the Pivotal Probability step, it is clear that o does indeed fulfill the first requirement. As for 'smooth invertibility,' o can take on either value, o_1 or o_2, no matter which outcome results. For each value of o, the random variable 'heads' – 'tails' is 1-1 with possible θ values. Finally, the differentiability assumption is not needed since the probabilities, for the discrete case, are well defined. Thus the discrete example is a bona fide illustration of the fiducial argument, at least given the reconstruction so far.

Next, review the bounded uniform distribution case to insure that the 'smooth invertibility' specification blocks the use of $\mathbf{F}$, the cumulative distribution function, as a 'natural' pivotal. $x_{\mathbf{F}}$, which is uniformly distributed on $[0, 1]$, is equal to x/θ. Even after fixing an upper bound for θ, namely $\theta \leqslant 15$, and prior to knowing x, $x_{\mathbf{F}}$ is capable of assuming any value in $[0, 1]$.

However, after observing $x = 10$ and from the specification of the stochastic distribution for x, it follows that $10 \leqslant \theta \leqslant 15$, so that $\frac{2}{3} \leqslant x_F \leqslant 1$. But this restriction on the range of F, the pivotal, is in violation of (A) of the smooth invertibility requirement. Therefore, x_F is not 'natural' and the fiducial step is blocked. There is a procedure for salvaging a modified fiducial argument for this example. In order to present it, an intermediate development must be attended to first.

4.3. ANCILLARITY, SMOOTH INVERTIBILITY AND THE CC PRINCIPLE

To this point, the reconstruction has assumed the existence of a sufficient statistic; a completely adequate reduction of the initial experimental data, which is of the same dimension as the unknown parameter. (For the discrete version of the fiducial argument, comparable dimensionality between statistic and parameter is replaced by a requirement that the parameter space and sample space for the sufficient statistic be 1-1.)

One of the last improvements of the fiducial argument which Fisher considered was the use of an ancillary statistic to generate an exhaustive statistic in place of a (missing) sufficient statistic. This maneuver can be incorporated into the argument by suitably weakening the 'smooth invertibility' requirement (A). To begin, re-examine several pertinent definitions (suppressing BK).

DEFINITION: A reduction of the data x to $t'(x)$ is a *sufficient statistic* for θ iff:

$$p(x/\theta) = p(x/t'(x)) \cdot p(t'(x)/\theta);$$

where the first factor to the right of the equality sign is independent of θ.

DEFINITION: A reduction of the data x to $t''(x)$ is an *ancillary* statistic for θ iff:

$$p(x/\theta) = p(x/\theta \,\&\, t''(x)) \cdot p(t''(x));$$

where $p(t''(x))$ is independent of θ.

DEFINITION: A statistic is *exhaustive* for θ if it becomes sufficient for θ when conditionalized on some ancillary statistic for θ.

As before, two examples are given: one continuous (taken from *Statistical*

Methods and Scientific Inference[10]) and the other discrete, a modification of Hacking's example.

Continuous

BK: The pair of observables (x, y) are distributed in a bivariate distribution $dF = \exp[-(\theta x + y/\theta)]\,dxdy$ with θ unknown, $x, y > 0$.

Pivotal: Define statistics t and u by: $t = \sqrt{y/x}$ $u = \sqrt{xy}$. t is the maximum likelihood estimate for θ, the best reduction of the data, but t is *not* sufficient. Let pivotal $\tau = \ln t - \ln \theta$.

Pivotal probability: Given u, τ is distributed
$[\frac{1}{2}K_0(2u^2)] \exp[-2u \cdot \cosh \tau]\,d\tau$;
where K_0 is the Bessel function, independent of θ.

Exhaustive step: u is an ancillary statistic for θ since its distribution is independent of θ, given by $dF = 4K_0(2u)du$. The sampling distribution of t, conditional on u, is $[\frac{1}{2}K_0(2u)] \cdot \exp[-u(t/\theta + \theta/t)] \cdot dt/t$.

Fiducial step: If τ retains its conditional distribution, given u, after observing t, then the integral of the pivotal probability gives the probability that θ is between $t \cdot \exp[\tau_1]$ and $t \cdot \exp[\tau_2]$.

Discrete

BK: A binomial coin ($\theta = .4$ or $\theta = .6$) is flipped *twice* with the random variable (x, y), each flip having two possible outcomes o_1 or o_2 ('heads'-'tails'). The trials are independent with $p(o_1/\theta = .6) = p(o_2/\theta = .4) = .6$

Pivotal: Define a pivotal v by:
v_1 occurs iff $(\theta = .6$ & $[o_1, o_1])$ or $(\theta = .4$ & $[o_2, o_2])$.
v_2 occurs iff $(\theta = .4$ & $[o_1, o_1])$ or $(\theta = .6$ & $[o_2, o_2])$.
v_3 occurs iff $(\theta = .6$ & $[o_1, o_2])$ or $(\theta = .4$ & $[o_2, o_1])$.
v_4 occurs iff $(\theta = .4$ & $[o_1, o_2])$ or $(\theta = .6$ & $[o2, o_1])$.

Pivotal probability: v is distributed independently of θ, with $p(v_1) = .36$, $p(v_2) = .16$, $p(v_3) = p(v_4) = .24$.

Exhaustive step: Define the statistic $t(x, y)$ by: $t(x, y) = \textit{same}$, iff $x = y$, i.e. $[o_1, o_1]$ or $[o_2, o_2]$; $t(x, y) = \textit{opposite}$, iff $x \neq y$, i.e. $[o_1, o_2]$ or $[o_2, o_1]$. Note that $p(\textit{same}) = .52$, and $p(\textit{opposite}) = .48$. Thus, $t(x, y)$ had a distribution independent of θ and is ancillary.

Fiducial step: If v retains its conditional distribution, given $t(x, y)$, then it yields a fiducial probability for θ when (x, y) is observed. The distribution for v, given $t(x, y)$, is obtained by normalizing by the value $1/p(t(x, y))$.

How does the exhaustive step work? What is its connection to (*A*) of smooth invertibility? Consider the second question first. By examining the discrete case, it is clear that, without use of $t(x, y)$, v is not a smoothly invertible pivotal. That is, prior to observing (x, y) v can take on any of four values. However, having seen (x, y) v is restricted to two values only; either v_1 and v_2, or else v_3 and v_4. By conditionalizing on $t(x, y)$, v is restricted

to only a pair of values: just the pair left after seeing (x, y). Thus, if conditionalizing on an ancillary statistic in order to bypass (A) of smooth invertibility is permitted, the fiducial argument can be extended to include exhaustive statistics. To repeat, the revision depends on a weakening of the first constraint of the 'smooth invertibility' requirement.

The question now arises, what justification is there for the use of ancillarity, through exhaustive estimation, to modify the recognizability conditions? One answer will not do. It is *not* the case that conditionalizing on the ancillary statistic leaves probability statements about the pivotal invariant. Thus, restricting the reference class for statements about the pivotal by holding the ancillary statistic constant yields a *recognizable* subset. For example, the probability that $v = v_1$, prior to knowing $t(x, y)$, is .36, but after observing $t(x, y)$ it is either 0 or $.36/p(t(x, y)) = .36/.52$.

There is an available defense, however, which rests on two principles: Confirmational Conditionalization, and the claim that ancillarity is a derived irrelevance condition, irrelevant to the unknown parameter. Confirmational Conditionalization [CC] is defined as follows (see p. 7):

> [CC] If E and E' represent a pair of mutually consistent sentences and $e \,\&\, E = E'$, then if the conditional probability for h given $e \,\&\, E$ equals r', it follows that $p(h/E') = r'$.

The claim that ancillarity is tantamount to irrelevance is defended by appealing the Bayes' theorem and CC. This use of Bayes' theorem is not to be construed as an admission of specific 'priors.' Quite to the contrary, the argument given here is that if ancillarity is a legitimate irrelevance condition, then it must be so generally; that is, for all possible prior probability assignments. In other words, the calculus of probabilities along with Bayes' theorem is to be used as the check or audit of the probabilistic accounting system wherein the irrelevance claim is made. The underlying explication of inductive relevance is, of course, just the familiar Carnapian or Bayesian one; to wit, given the background knowledge BK, e is relevant to h iff $p(h/BK) \neq p(h/BK \,\&\, e)$.

For emphasis, the Bayesian argument for ancillarity is again rehearsed (see p. 20). From Bayes' theorem and CC it follows that (suppressing BK):

$$p(\theta/x) = \kappa \cdot p(x/\theta) \cdot p(\theta).$$

Let $t(x)$ be ancillary for θ, i.e. the distribution for $t(x)$ is independent of θ. Then

$$p(\theta/x) = \kappa \cdot p(x/\theta \,\&\, t(x)) \cdot p(t(x)).$$

But, since $\mathbf{p}(\mathbf{t}(x))$ is independent of θ, it can be absorbed into the constant κ. Hence, the correction for conditioning on $\mathbf{t}(x)$ is $1/\mathbf{p}(\mathbf{t}(x))$, and $\mathbf{t}(x)$ is irrelevant for θ.

By way of illustration, reconsider the continuous example of p. 113. The joint density for t and u, given θ, is just:

$$d\mathbf{F} = 2 \cdot \exp\left[-u(t/\theta + \theta/t\right] \cdot u \cdot dt/t. \qquad (*)$$

The marginal distribution for u, the ancillary statistic for θ, is given in the information presented in the Exhaustive step on p. 113. Dividing (*) by $4K_0(2u)$ (from the marginal density for u), gives (to a constant) the conditional density of the (now) exhaustive t statistic. This is stated in the last line of the Exhaustive step.

The upshot of this argument for exhaustiveness, through CC and Bayes' theorem, demonstrates the incompleteness of Fisher's 'recognizability' condition (for the pivotal variable) as the explanation for the Fiducial step. To repeat, the conditioning maneuver for exhaustiveness violates the 'recognizability' restriction with respect to the pivotal variable. However, the desired irrelevance claim can be maintained by appeal to CC and Bayes' theorem. Conditioning on the ancillary statistic gives irrelevant information about θ (*not* about the pivotal variable), which is, of course, the sort of irrelevance needed for the Fiducial step.

Using confirmational conditionalization, without use of ancillarity, permits the loosening of (A) of the smooth invertibility requirement (see p. 111) in yet another way. Turning to the bounded uniform distribution problem of p. 110, we see that if the information about the upper bound, call it d: that $\theta \leqslant 15$, is suppressed for the moment, then $x_{\mathbf{F}}$ meets all the conditions for a 'natural' pivotal. Observing x, whatever it is, induces a fiducial distribution for θ. Now, using CC, by adding d to the modified background knowledge, i.e. $BK \,\&\, x$, to reach $(BK \,\&\, x \,\&\, d)$ the final evidential basis, a conditional fiducial distribution is made available. That is, using CC, the extra information about the upper bound (d) can be incorporated into the background knowledge $BK \,\&\, x$, *after* the fiducial argument has been used to generate a posterior probability of θ based on $BK \,\&\, x$.

This mode of inference has, as its ultimate justification, the mathematical fact that Bayes' theorem is symmetric about the order of additional evidence, using CC. That is, it does not matter in which order the evidence is introduced.

$$\mathbf{p}(\theta/BK \,\&\, e_1 \,\&\, e_2) \propto \mathbf{p}(e_1/BK \,\&\, \theta \,\&\, e_2) \cdot \mathbf{p}(e_2/BK \,\&\, \theta)$$

$$= \mathbf{p}(e_1 \,\&\, e_2/BK \,\&\, \theta)$$

$$\propto p(e_2/BK \,\&\, \theta \,\&\, e_1) \cdot p(e_1/BK \,\&\, \theta)$$

$$= p(e_1 \,\&\, e_2/BK \,\&\, \theta).$$

Of course, for the fiducial argument, the order is all important. This conclusion can be supported without pain of contradicting the probability calculus by noting that the fiducial argument is intended to apply only in special 'ignorance' contexts. If the background knowledge includes enough data to permit a frequency based prior probability for the unknown parameter (by a direct inference), then on that evidential basis the fiducial argument does not apply.

For the case in point, the TRE condition prohibits deriving a fiducial argument when datum d: $\theta \leqslant 15$ is known. This prohibition is enforced by condition (A) of smooth invertibility. However, *without* violating TRE, the investigator can 'build up' to his current knowledge state, using CC and Bayes' theorem, so that the end state is probability based on all the evidence: $BK \,\&\, x \,\&\, d$. The importance of this *sequential method* is that at one stage, e.g. given just $BK \,\&\, x$, the fiducial argument applies.

Having opened the door to the use of CC and Bayes' theorem for combining evidence probabilistically, we are left with the task of analyzing the three cases or possible combinations involving the fiducial argument (for a single parameter distribution) to determine that coherence is maintained. These cases are the combination of two bits of evidence: (i) where only one is capable of inducing (on its own) a fiducial distribution, i.e. data of two kinds; (ii) where both data bits are capable of inducing (on their own) a fiducial distribution and, moreover, can be combined into a single sufficient (or exhaustive) statistic capable of generating a fiducial probability; and (iii) the variation of case (ii) where there is no common sufficient (or exhaustive) statistic. Combinations of more than two groups of data can be treated in pairs; hence, can be reduced to one of the previous three cases.

4.4. CASE 1: THE FIDUCIAL ARGUMENT WITH DATA OF TWO KINDS

The bounded uniform distribution example, discussed above, is a simple instance of the first combination set. The two sorts of data, x and d, are different in that x (in the absence of d) is suitable for generating a fiducial probability. However, d is not suitable evidence for a fiducial argument. The combined evidence is treated by first using x to get a fiducial probability for θ, and then using CC (and Bayes') to properly modulate the first

probability by the addition of d. Fisher was much concerned with the problem of combining evidence subject to the probability calculus, and discussed this very case in an important article 'Some Examples of Bayes' Method of the Experimental Determination of Probabilities *A Priori*,' and also in his last text *Statistical Methods and Scientific Inference*. The procedure he suggests is identical to the one sketched above. A worthwhile exercise is to review an elementary problem, treated in his article and book.[11]

Imagine an experiment which consists of taking two kinds of readings on a radio-active source, the problem being to determine a probability for the unknown rate of decay, θ. One device, called machine W, is capable of continuous measurements, and records the time interval x, for a single emission. The other, detector O, registers for each of $a + b = n$ time intervals of ξ seconds (each) whether or not a particle (at least one) has been received.

The background knowledge is rich enough to determine that the c.d.f. $F(x, \theta) = 1 - \exp[-x\theta]$ for the unknown θ. The fiducial argument can be applied in this circumstance, for there exists a pivotal variable $w = x\theta$, with distribution independent of θ ($F(w) = 1 - \exp[-w]$), and which satisfies the 'smooth invertibility' condition. Hence, the fiducial density for θ, given x, is $-\partial F/\partial\theta$, or $x \cdot \exp[-x\theta]$.

Next, define λ as $\lambda = x/\xi$. Then the probability of getting a 'no particle' count in a single ξ interval is just $\exp[-\xi\theta]$, denoted ρ, and the fiducial density for ρ is (from $-\partial F/\partial\theta$) $\lambda\rho^{\lambda-1}d\rho$. Thus the fiducial argument supplies a 'prior' probability for the outcomes of the second experiment with detector O. If a is the number of instances (out of n) when O registered 'no particles,' then a has a simple binomial distribution, i.e. $[(a+b)!/a!b!] \cdot \rho^a \cdot (1-\rho)^b$. Thus, the first observation yields a fiducial probability which can be combined through Bayes' theorem with the second observation. It is important to note that the roles cannot be reversed, since the O outcomes are incapable of supporting a fiducial argument. The composite posterior distribution for ρ (and hence θ) is given by: $[(a+b+\lambda)!/(a+\lambda-1)!b!] \cdot \rho^{a+\lambda-1} \cdot (1-\rho)^b \cdot d\rho$. This concludes Fisher's example.

It should be clear from the foregoing analysis that Fisher was prepared to use fiducial probabilities just as if they were ordinary ones, derived by usual frequency principles. The technique employed in the 'radio-active' example makes no special use of the genesis of the 'prior' probabilitity. It might just as well have been supplied by a stochastic distribution, instead of the fiducial argument. It seems appropriate, therefore, before continuing with the other two single-parameter cases, to review the commitments already made (and the list is formidable), to determine that fiducial probabilities are ordinary.

The first difficulty to present itself (a variant of a paradox due to Edwards) arises from the use of ancillarity to extend the fiducial argument to cover exhaustive statistics.[12] Reexamine the discrete example of p. 113. Notice that if $t(x, y)$ equals *opposite*, then no matter which observation occurs, i.e. for both $[o_1, o_2]$ and $[o_2, o_1]$, the posterior fiducial probability for θ is identical. Specifically, conditionalizing on $t(x, y) = opposite$, $p(v_3) = p(v_4) = .24/.48 = .5$. Therefore, $p(\theta = .6/[o_1, o_2]) = p(\theta = .4/[o_1, o_2]) = p(\theta = .6/[o_2, o_1]) = p(\theta = .4/[o_2, o_1]) = .5$. But then given $t(x, y) = opposite$, it is not necessary to know exactly which outcome resulted, since they all lead to the same posterior probability. Intuitively, given $t(x, y) = opposite$, the information of the particular outcome is irrelevant. Also, since $t(x, y)$ is ancillary for θ, and ancillarity was defended as an irrelevance condition, it would appear that from two bits of irrelevant evidence, $t(x, y)$ and the particular outcome (given $t(x, y)$), the fiducial argument has been able to extract a probability for θ where none was available before. Certainly fiducial probabilities are not ordinary frequentist ones if from irrelevant evidence alone, and without use of unwarranted priors, they are so easily produced.

The response to this charge is two-fold: a concession and a subsequent defense (based on A. Birnbaum's analysis).[13] First the concession. As with simple tests of significance, those fiducial arguments (presented thus far) which do not rely on the exhaustive step are capable of a direct frequency interpretation, produced from repeated trials on the same experimental set-up. Throughout the reconstruction this potential interpretation has been noted, and consistently rejected: it is an improper one. For significance tests the correct interpretation was by means of the disjunction: either a rare event occurred or the null hypothesis is false. For the fiducial argument the correct interpretation is through the reference class for the pivotal. By conditioning on the ancillary statistic, the pivotal is divided into 'sub'-pivotals, and the normalizing procedure destroys any frequency interpretation for the original pivotal, based on repeated sampling of the experiment. There is, of course, a round-about repeated sampling interpretation, derived from the conditioning step of the ancillary statistic. For example, by repeating the pair of flips and conditioning on those with $t(x, y) = opposite$, 50% will result in v_3 and 50% in v_4. But the proportions in the overall sequence of repeated trials will be 24% each. Thus, fiducial probabilities obtained through the exhaustive step are not ordinary, in that they cannot be given the same interpretation as those derived without the use of ancillarity. To repeat, this sort of

interpretation is not the appropriate one anyway, so, from the Fisherian perspective, nothing is lost by the concession.

The other facet of this problem receives an illuminating explanation, based on some well known analysis, due to A. Birnbaum. He has shown that the conjunction of Sufficiency with Ancillarity is logically equivalent to the Likelihood Principle of irrelevance. All three rules are derivative on the combination of CC with Bayes' theorem (plus, of course, the standard explication of relevance). Thus, by adopting the Likelihood Principle, or the equivalent formulation through Sufficiency and Ancillarity, the fiducial argument follows rules compatible with Bayesian inference. Birnbaum has also shown (as has this analysis) that Neyman-Pearson confidence, or more precisely confidence analyzed through repeated sampling, is inconsistent with the Bayesian program. Therefore, the parting of the ways for fiducial and confidence approaches can be traced to the use, by fiducial techniques, of these two principles. It follows (from Birnbaum's work) that Likelihood and Neyman-Pearson confidence are incompatible. Therefore, whenever the fiducial argument includes the Exhaustive step, which means admitting ancillarity, the confidence (repeated sampling) interpretation is no longer guaranteed. Of course, the use of Sufficiency is common to both programs (as has been discussed before), so that it is Ancillarity which causes the split.

Examine the discrete example from the perspective of Likelihood. The Likelihood Irrelevance Principle [LIP] states that if two bodies of evidence each supply the same likelihood function (up to a constant) then they are of equal evidential value for inference about the unknown parameter. A special instance of LIP is that if evidence $d \,\&\, e$ yields the same likelihood function (for θ) as does d alone, then e adds no relevant evidence for inference about θ. Given $t(x, y)$ equals *opposite*, the likelihood for $\theta = .4$ is the same as for $\theta = .6$, namely .48, so that the likelihood ratio (ignoring constants) is 1. The likelihood for $\theta = .4$, given the actual experimental outcome, i.e. $[o_1, o_2]$ or $[o_2, o_1]$, equals that for $\theta = .6$, both being .24. Ignoring constants, we see that the likelihood ratio is again 1. Therefore, by the special instance of LIP, the extra information gained, once $t(x, y)$ is known to be *opposite*, is irrelevant.

$t(x, y)$, the ancillary statistic, is itself irrelevant for inferences about θ, but to establish that fact one must make use of the combined CC and Bayes' rules. Review the argument on p. 114. We showed there that the use of Bayes' theorem for establishing the validity of ancillarity does not commit the investigator to specific prior probabilities. Therefore, the conclusion that each of $[o_1, o_2]$ and $[o_2, o_1]$ is irrelevant for inferences about θ does not

depend on a particular prior probability for θ. However, since the fiducial argument supplies a precise value for $\mathbf{p}(\theta/[o_1, o_2] \,\&\, BK)$ and equally for $\mathbf{p}(\theta/[o_2, o_1] \,\&\, BK)$, it follows that these fiducial posterior probability values must equal the masked prior probability value for θ, or else the data would be relevant.

Specifically, using the fiducial argument, $\mathbf{p}(\theta = .4/\text{either outcome} \,\&\, BK) = \mathbf{p}(\theta = .6/\text{either outcome} \,\&\, BK) = .5$. Hence the masked, conjugate prior probability is: $\mathbf{p}(\theta = .4/BK) = \mathbf{p}(\theta = .6/BK) = \frac{1}{2}$. Thus the fiducial argument requires a singular conjugate prior probability factor if it is to be coherent. The values $(\frac{1}{2}, \frac{1}{2})$, the conjugate fiducial prior probability, looks suspiciously like the Objective Bayesian ignorance or indifference prior. That this is not an accidental coincidence will be made clearer in the next section, where D.V. Lindley's work (essential to the discussion of the second of the three cases) is examined.

4.5. CASE 2: THE FIDUCIAL ARGUMENT WITH DATA OF ONE KIND AND A COMMON SUFFICIENT STATISTIC

The first case treated composite evidence where just one of the two evidential bits was capable of supporting a fiducial argument. The rest of the evidence was accounted for only after the first had been used to generate a fiducial 'prior.' The second case to be considered is where both of the evidential parts are suitable fiducial material, and also where they can be combined into a single sufficient (or exhaustive) statistic, which then can be used to produce a fiducial probability.

An example of this is already available to the reader. Look yet again at the discrete binomial example of p. 113, and compare it to the first discrete (Hacking's) example of p. 108. Notice that either flip, i.e. either x or y, is capable of generating a fiducial probability for the unknown θ. However, the technique actually employed with the joint evidence (x, y) uses a simultaneous exhaustive step, instead of a sequential one as in the preceding section. The question now arises, what happens if the sequential process is used in place of the simultaneous one? Will the two posterior probabilities agree?

For the discrete example the answer is yes. That is, one can construct a posterior probability for θ based on both observations in either manner without a conflict. The simultaneous technique has already been thoroughly discussed for this problem. By way of further illustration of the sequential maneuver, consider the following analysis.

The sequential method offers two possibilities. Remembering that Bayes'

theorem is symmetric about the order of the observations, and using CC, one can construct the fiducial 'prior' from either x or y, and use the other analogous to data from detector O to modulate the first probability. If $t(x, y)$ equals *same*, then the selection of x or y to serve as the fiducial data (analogous to data from W) makes no difference. In this case the fiducial 'prior' for θ is (by Hacking's example) either .6 (for $\theta = .6$ given 'heads,' or for $\theta = .4$ given 'tails') or .4 (otherwise). The second observation, assumed to be the same as the first, can be combined with the first using Bayes' theorem and CC to yield the overall posterior given as:

$$p(\theta = .6/[o_1, o_1]) = .6 \cdot .6 \div (.36 + .16) = .36/.52$$
$$p(\theta = .4/[o_1, o_1]) = .4 \cdot .4 \div (.36 + .16) = .16/.52$$
$$p(\theta = .6/[o_2, o_2]) = .4 \cdot .4 \div (.36 + .16) = .16/.52$$
$$p(\theta = .4/[o_2, o_2]) = .6 \cdot .6 \div (.36 + .16) = .36/.52$$

All of these figures are in exact agreement with those obtained from the simultaneous, exhaustive method.[14]

Before considering Lindley's general analysis of the consistency of the two methods, it is interesting to note Fisher's response. In reply to Lindley's papers, Fisher wrote in 1959, just three years before his death,

> In 1956 the author developed a method of combining data of two distinct kinds, one portion being of a type capable of leading to an exact probability distribution by a fiducial argument, while the remainder, taken by itself, like the Bernouillian type of data considered by Bayes, leads to no statement of probability whatever. The procedure was essentially that of Bayes himself, using the first portion to obtain probabilities *a priori* while the second part, providing likelihood values only, supplies the multipliers by which the probabilities *a priori* are converted into logically similar probabilities *a posteriori*.
>
> It had not occurred to me to consider whether this method might be applicable to the Combination of Observations in a more general sense, partly because the combination of observations of a similar kind presents no new problem (it must introduce the question of exhaustive estimation for the data as a whole, and not merely for one of its parts), partly also because Bayes himself did not discuss possible applications in this wider field.[15]

After introducing his discussion with this explanation and excuse, Fisher goes on to show that an example proposed by Lindley,[16] in illustration of the potential *incompatibility* of the two methods is actually in error; so that the two procedures in Lindley's example are consistent. Specifically, it involves the combination of data from two samples of a normal distribution,

with unknown variance and mean, i.e. the combination of two 'Student's' t-statistics, for inference about the unknown mean. Lindley's error is a technical one and, since the problem is multivariate, the discussion of his example will await the reconstruction for the multivariate (step-by-step) fiducial argument.

The relevant point of Fisher's response is that, after rebutting Lindley's criticism he goes on to say that,

> Mr. Lindley writes as though he had established important new limitations to the use of Bayes' method, but it should be remembered that he is not investigating the combination of data of two different kinds as carried out by Bayes, and discussed in my book. They are not in any sense limitations of the fiducial argument, but only of Lindley's proposed extension of the use of Bayes' argument.[17]

Now there is a certain peculiarity to Fisher's method of argumentation (which might explain some of the indignation on the part of his critics). The limitations to which Fisher refers appear in a separate paper by Lindley.[18] They will be analyzed shortly. Fisher recognizes the validity of Lindley's concern over the consistency of the two fiducial methods (or else why bother to point out Lindley's error in the first paper), but then he withdraws from the debate when it becomes clear that Lindley has discovered new paradoxes for fiducial inference. For the sake of *his* reconstructed Fiducial method, Fisher chooses to put the blame on Bayes' argument, suggesting that it might be restricted to observations of two kinds only.[19]

Bayes' 'argument,' as Fisher chooses to call it, relies on CC, Bayes' theorem and, of course, the existence of suitable 'prior' probabilities. There is nothing inherent to Bayes' method which restricts its applicability to observations of two kinds only. Fisher may be correct, historically, when he argues that Bayes' example uses observations of two kinds; but the question of validity of the method is not an historical one, as Lindley so rightly points out.

Reviewing the situation, one sees that each of the prerequisites for Bayes' argument are met for the problem at hand. That is, CC is to be admitted, as before; Bayes' theorem retains its validity, as always; and the problem of the existence of a frequency based prior is answered by using the fiducial argument on one of the data bits, just as in the sequential method with observations of two kinds. If Lindley has discovered some limitations to the consistent application of this method (with the simultaneous one), *and he has*, then they must be faced up to, Fisher's protestations not withstanding.

Lindley's argument is both direct and elegant. Assuming what Fisher states, namely that the Fiducial probability is given as $-\partial F/\partial\theta$ when it exists, and granting the premises of the case, i.e. that the two observations are

each capable on their own of inducing a fiducial distribution (subject to the two pivotal constraints), and are capable of being combined into a single sufficient statistic appropriate for a fiducial argument (of the simultaneous sort), then the two fiducial methods are mutually consistent if and only if the unknown parameter is a location (or scale) parameter for the observations (or transformable thereto) with a uniform 'improper' prior probability as the conjugate one. Actually, Lindley proves a stronger result as well. The necessary and sufficient conditions for consistency of the two methods are also the necessary and sufficient conditions for consistency of the fiducial argument with CC, Bayes' and the existence of priors, even for a single observation. In other words, for any one parameter fiducial distribution to be coherent it must meet the conditions set forth by Lindley.

The two results raise two objections. Using the stronger conclusion first, it might be argued that however the fiducial argument is patched-up to be made coherent, its use merely hides the underlying Bayesian commitments, particularly the uniform 'improper' prior probability. Less politely, it could be argued that fiducialists who try and stay coherent are really closet Bayesians, who also believe in 'ignorance' priors: to wit, the uniform 'improper' ones.

The Fisherian response is not hard to come by. True, fiducial probabilities are coherent just in case Lindley's conditions hold; but the existence of compatible conjugate priors, even of a special 'ignorance' sort, do not establish their legitimacy as frequency based priors. CC, the apparent focal point of the Bayesian argument, commits the fiducialist to priors only by being run backwards; from the epistemic state $BK \,\&\, x$, where the fiducial argument applies, to BK alone – and then seeing what Bayes' theorem demands as a suitable starting point. It is at this starting place that Fisher insists that no prior probability exists at all.[20]

As desirable as such a conclusion must be to a true fiducialist, it misses the mark. Admittedly, the conjugate priors cannot be derived by the Frequency Principle in a direct inference. If they could, then the fiducial argument would not apply – the epistemic state being too 'rich' to support the ignorance claim. However, the fiducial argument is intended as a solution to inverse inference in cases of ignorance, and if the solution is satisfactory then it must be coherent (at least). Coherence entails a Bayesian 'model' within the probability calculus, equipped with all the trappings, including a conjugate prior. It makes little sense to cry foul when the conjugate priors are exposed by Lindley's argument, on the one hand; yet to benefit from the existence of the Bayesian model, ensuring coherence, on the other.

Perhaps the fairest statement of the conclusion forced by Lindley's examination is that, although fiducialists *are* committed to the conjugate priors, the commitment does not arise from a traditional Objective Bayesian principle: expressed as a rule for extracting prior probabilities through symmetries (or claims of subjective indifference) over the unknown parameter(s). In the place of such principles, the fiducialist uses regulations based on the stochastic properties of the random variable – which have the advantage of being frequency based. Thus the observer invariant objectivity of the fiducial argument stems from a lack of essentially subjective features, apart from the specification of the back-ground evidence. Hence, the resulting conjugate priors are permissible, from a frequentist standpoint, if the fiducial argument is an approved inference.

4.6. CASE 2: LINDLEY'S PARADOX AND CANONICAL PIVOTAL VARIABLES

The second objection is much more serious. Using Lindley's results about the necessary and sufficient conditions for consistency of the two fiducial methods, it can be argued that Fisher's restraints on the pivotal are *not* adequate for guaranteeing the desired congruity. Note that this objection does not rely on a Bayesian ignorance prior, for the consistency question uses only fiducially derived probabilities for priors, where needed. Lindley brings the point home with an illustration.[21]

Let x be the observed random variable. Define $F(x, \theta)$, the c.d.f., by its density function $\partial F/\partial x = f(x/\theta) = [\theta^2/(\theta+1)] \cdot (x+1) \cdot \exp[-x\theta]$, in violation of the requirement that F be transformable to location parameter form. If (x_1, x_2) are two independent observations from this distribution, then $z = x_1 + x_2$ is a sufficient statistic for θ, based on the joint (x_1, x_2) data. The cumulative distribution function for z is a smoothly invertible pivotal, so that it induces a fiducial distribution for θ, without appeal to the exhaustive step. The alternative sequential method, constructed by using either of (x_1, x_2) to induce a fiducial probability for θ and then incorporating the other observation by use of Bayes' and CC, leads to an incompatible posterior probability. In fact, by reversing the roles of x_1 and x_2 in the sequential process (and this can be done from the initial assumption that both observations are of the same kind) a third posterior probability is created, inconsistent with either of the other two. This last result further muddies the waters, since Bayes' theorem is symmetric about the order of the evidence. Hence, there is conflict between the two sequentially derived fiducial probabilities, without worrying about the third conflict: the

inconsistency with the fiducial probability derived from the single sufficient statistic z.

As intolerable as this state of affairs is, it would be still less desirable to rest with Fisher's response, that Bayes' method should be restricted to observations of two kinds exclusively. The position adopted here is that the 'smooth invertibility' condition is an insufficient restraint and that it does not explicate Fisher's 'recognizability' condition. Just as the initial condition (1), the existence of the pivotal, proved to be inadequate and was supplemented with (2), smooth invertibility, Lindley has shown that the conjunction of (1) and (2) is also incomplete and must be amended some more. To evaluate the new criterion, three conditions are taken to be necessary. First, the difficulties arising from the Lindley analysis must be avoided. Where alternative fiducial methods are available they must all yield mutually compatible probabilities. Second, the revised conditions must be in the frequentist spirit of Fisher's proposals. There can be no appeal to unwarranted prior probabilities. Third, a further connection to the rest of the Fisherian program is required. Despite the obviously strained faith in Fisher's ability to identify necessary and sufficient conditions for fiducial inference, let us retain requirements (1) and (2) and add a third. For the lack of a better phrase the new restraint on pivotals limits fiducial inference to *canonical* pivotals.

DEFINITION: A pivotal variable v is *canonical* iff the distribution for v is the same as the distribution for x (the random variable to be observed) when x is substituted for v and $\theta = \theta^*$, for some θ^* in the parameter space.

Before checking the additional qualification against the three conditions of adequacy, it is best to review several examples of canonical pivotals.

Example$_1$: Let $\mathrm{F}(x, \theta) = 1 - \exp[-x\theta]$; $x, \theta > 0$. That is, the exponential distribution, found in Fisher's 'radio-active' problem. Let $v_1 = x\theta$. Then v_1 is a canonical pivotal since the distribution for v_1 is exponential with $\theta^* = 1$. That is, $\mathrm{F}(v_1) = 1 - \exp[-v_1]$, which is independent of the unknown θ. It has already been shown that v_1 is 'smoothly invertible.'

Example$_2$: Let $\mathrm{F}(x, \theta) = x/\theta$; $\theta > 0$, $0 \leqslant x \leqslant \theta$. That is the uniform distribution, discussed on p. 110. Let $v_2 = x/\theta$, the value of the cumulative distribution function itself. Then v_2 is uniformly distributed on $[0, 1]$, and is a canonical pivotal with $\theta^* = 1$. It has already been shown that v_2 is 'smoothly invertible.'

Example$_3$: Let $\mathrm{F}(x, \theta) = \mathrm{N}(\theta, 1)$; x, θ unbounded. That is the normal

distribution with unit variance, discussed on p. 108. Let $v_3 = x - \theta$. Then v_3 is distributed as N(0, 1), which is independent of the unknown θ, and is canonical with $\theta^* = 0$. It has been shown 'smoothly invertible.'

Example$_4$: Let $F(x, \theta)$ be the discrete binomial distribution as given on p. 108, i.e. Hacking's example; $\theta = .4$ or $.6$, $x =$ 'heads' or 'tails.' Let v_4 be defined as:

$$v_4' \text{ iff } (\theta = .6 \ \& \text{ 'heads'}) \text{ or } (\theta = .4 \ \& \text{ 'tails'})$$
$$v_4'' \text{ iff } (\theta = .4 \ \& \text{ 'heads'}) \text{ or } (\theta = .6 \ \& \text{ 'tails'}).$$

Then v_4 has the binomial distributions $p(v_4') = .6$, $p(v_4'') = .4$, which is independent of the unknown θ, and is canonical with θ^* equal to either .4 or .6. It, too, has already been shown to be 'smoothly invertible.'

Thus, with the exception of Lindley's paradoxical distribution, all smoothly invertible pivotals presented so far are also canonical pivotals. Therefore, it is not necessary to update the previous analysis for those distributions.

How does the third constraint serve to avoid Lindley's counterexample and, more generally, the threat of inconsistency of the fiducial argument stemming from alternative methods? The answer is that by restricting pivotals to those which are smoothly invertible and canonical, one satisfies the necessary and sufficient conditions for coherence (as proposed by Lindley). (The proof of this claim is found in Appendix 9.3.) Hence, since his counterexample leads to a failure of coherence, there is no pivotal which is smoothly-invertible and canonical for that distribution. The first requirement for a satisfactory reconstruction has been met. The conjunction of (1) the existence of the pivotal variable with (2) the smooth invertibility of the pivotal variable and (3) the canonicity of the pivotal variable secure the coherence of fiducial inference by permitting a reduction to Lindley's restrictions.

The second condition of adequacy, that added demands do not introduce unwarranted prior probabilities, is satisfied by the formulation of canonicity. Canonicity limits the pivotal variable's distribution by specifying its relation to the distribution model for the variable to be observed. There is no appeal to any prior probability for the unknown parameter.

The third condition of adequacy, that there be a tie to the rest of the Fisherian program, is more of a challenge. The response offered here attempts to link fiducial inference to significance testing (and indirectly to likelihood). One of the key steps in the proof of the claim that smoothly invertible, canonical pivotals induce coherent fiducial probabilities (see Appendix 9.3) is the result:

$$|\partial V/\partial x| = 1, \text{evaluated at } \theta = \theta^*.$$

That is, for the special θ value by which v, the pivotal, is canonical, $|\partial V/\partial x| = 1$. But this ratio is just the factor that is involved in transforming random variables. In other words, if one were to switch random variables from x (the observable) to v (the pivotal), this factor (the degenerate Jacobian) emerges. However, at $\theta = \theta^*$ it is unity, so there is no change in distributions. More precisely, at $\theta = \theta^*$ (and this value only) v and x are identical. Not only do they have the same distribution (as required by *canonicity*) but they are indistinguishable.

How can this fact be used to interlace significance testing with fiducial probability? Suppose one were to make a prediction about the pivotal, despite the fact that its value is never truly observed, since θ never is. The distribution for the pivotal is independent of the unknown parameter, and given just the background knowledge, BK, and using D_p to rank possible outcomes of v, a critical region for v, of size α, can be selected. Of course, since there is no null hypothesis being tested, no conditional prediction is necessary and step (2*b*) (p. 81) can be eliminated. Compare this prediction to that derived from the same D_p ordering and α level for the null hypothesis $\theta = \theta^*$. Since $x = v$ for $\theta = \theta^*$ the two predictions will be identical. That is, the unconditional prediction for the pivotal, now identical to x, the random variable which serves as the observable of the significance test for $h_0: \theta = \theta^*$, is compatible with the conditional prediction of the significance test, i.e. $v. \in \sim C'$. step (2*b*).

For instance, using Example$_3$, the normal distribution with unit variance, the special parameter value for which $v_3: (x - \theta)$ is canonical is $\theta = \theta^* = 0$. But then $v_3 = x$, as argued in the preceding paragraph. v_3 is distributed normally, N(0, 1), independent of the unknown parameter. Using D_p, and setting $\alpha = .05$, the prediction for v_3 is that it will fall within 2 (standard deviation) units of 0. This prediction is based on BK and D_p only: no additional null hypothesis is required. On the other hand, the corresponding null hypothesis, $h_0: \theta = \theta^* = 0$ yields, at the .05 level and using D_p (to rank the x outcomes), a critical region for x such that h_0 is rejected iff x falls more than 2 (standard deviation) units from 0. But at $\theta = \theta^*$, $v_3 = x$, so that the null hypothesis will be falsely rejected just in case the prediction for the pivotal is incorrect (given $\theta = \theta^*$).

The next step is to generalize this last conclusion by eliminating the final, parenthetical, clause. Suppose that the null hypothesis h_0': $\theta = \theta' \neq \theta^*$ is changed from h_0 (above). Assuming that the problem has been put into

'location parameter' form (which can be assured from Lindley's result, e.g. if the problem is in 'scale parameter' form transform back by taking logarithms of both x and θ), then x and v are no longer identical, even though the distribution for the pivotal has not changed. For example, using the normal distribution, if h_0': $\theta = c \neq 0$ is postulated, then $v_3 = x - c \neq x$ (assuming h_0'). The prediction for v_3 is unchanged, but that for x, when testing h_0', has. Using D_p, and letting $\alpha = .05$ (as before), the new critical region for x is defined as: $C' = \{x: |x - \theta| > 2$ (standard deviation) units$\}$. But the prediction about v_3, namely that it will not be more than 2 (standard deviation) units from 0, is compatible with the conditional prediction for x, given h_0'. Therefore, since c is general, the null hypothesis will be falsely rejected just in case the prediction for the pivotal (which is insensitive to the null hypothesis) is incorrect.

Thus, Fisher's original 1930 formulation of fiducial probability through fiducial limits, identical to the set of parameter values left unrejected by corresponding significance tests, has been vindicated through the new canonicity condition. That is, for the single parameter fiducial argument (with or without exhaustive procedural step) the restriction to smoothly invertible, canonical pivotals (assuming D_p discrepancy ranking and location parameter formulation) guarantees an agreement between Fisherian significance tests and fiducial probability interpreted as a probability for unrejected hypotheses. Moreover, under location parameter distribution form, D_p and likelihood agree (see p. 80) so that all three Fisherian measures of support: significance, likelihood and fiducial probability, become mutually consistent.

4.7. TWO NOTES ON THE CONNECTION BETWEEN SIGNIFICANCE AND FIDUCIAL PROBABILITY

There are two supplementary notes to be considered before continuing on to the final (of the three) cases for single parameter problems with composite data. The reader might be concerned with the need for location parameter form and the D_p discrepancy ranking. What is the connection?

As stated previously (see p. 80), D_p and likelihood agree for location parameter problems. That is, if D_p, the probabilistic discrepancy ranking (defined by using the conditional probability [density] $p(x/\theta)$ as a function of x) is compared to the likelihood ranking (defined using the conditional probability [density] $p(x/\theta)$ as a function of θ), then the two rankings agree for all possible outcomes x, and with respect to all possible parameter values

θ. In other words, the relative significance for a null hypothesis h_0 is directly linked to its relative likelihood.

The argument linking significance with fiducial probability, seen as fiducial 'intervals,' rests on the use of a common D ranking for making (i) predictions about the *r.v.* x, given h_0, and (ii) predictions about the pivotal variable v. By insisting on D_p as the discrepancy measure the resulting intervals are also likelihood ratio intervals and the significance tests are likelihood ratio tests. The agreement between the two predictions does not depend on the use of D_p, but on the use of *some* common D ranking. Besides the secondary benefits of using D_p, e.g. agreement with such N-P conclusions as the shortest confidence intervals or shortest unbiased intervals whenever likelihood ratio tests are favored, the way is paved for a connection with Hacking's reconstruction (which uses likelihood as the key criterion for irrelevance). Unlike the reconstruction presented here, Hacking's does not require any special distribution form (such as, location parameter or scale parameter distribution) as he appeals to the likelihood function, which is *invariant* to these sorts of transformations of the *r.v.* However, as will be pointed out, Hacking does require a special representation of the pivotal distribution (by 'experimental densities') in order that his likelihood method of support may be applied to hypotheses about the pivotal.

The second point to be considered is the reminder that the foregoing discussion, linking fiducial probability to significance, should not be abused by attempting to use the pivotal variable as a replacement observation for x; even when the null hypothesis is h_0: $\theta = \theta^*$. The temptation is to say that when $\theta = \theta^*$ the pivotal variable v is identical to x (which is the random variable to be observed); therefore, it seems natural to use v in place of x. To try this maneuver is to fall into the trap of misconstruing the role of the null hypothesis; that is, to repeat the Spielman-Redhead error of contaminating the observable with the hypothesis being tested. The transformation T: $x \to v$ may be both 1-1 and differentiable (using θ^* of h_0), but it remains epistemically incalculable so long as θ is unknown. This is all the more subtle for the special null hypothesis h_0, as then $v = x$, and the transformation is the degenerate identity map I: $x \to x$. But, of course, h_0 is not known and the special value θ^* cannot be used to transform the *r.v.*

4.8. CASE 3: DATA OF ONE KIND WITHOUT A COMMON SUFFICIENT STATISTIC, A PARADOX AND THE PRINCIPLE OF INSUFFICIENT REASON

The third and final single parameter, composite data, fiducial argument to be considered is one where each of the data bits are capable, on their own, of

inducing a fiducial distribution, but where no single sufficient statistic exists based on the composite data. That this case deserves special treatment is warranted by the facts that (1) based on the reconstruction thus far, there exists a paradox stemming from alternative methods for deriving fiducial probabilities, and (2) no satisfactory resolution is apparent, with the exception of reducing the paradox to an already existing *related* problem for which no general solution is known.

Before delving into the paradox, reexamine Fisher's remarks (p. 121) concerning the appropriateness of using Bayes' method (the sequential method) for combining two pieces of evidence and producing a single fiducial probability. In the second paragraph quoted there is a parenthetical note, which in effect discounts the current case as a legitimate one. That is, Fisher requires that the two data parts must be capable of being combined into a single summary, or else be of two kinds, i.e. the first case. If Fisher was content with his reply to the second case, that is, with the disclaimer of the sequential method when the data were of the same kind, then there would be no need for this last, additional requirement that the data be, as in case two, combinable into a single summary. The extra warning sounds the alarm for the danger looming ahead.

In Appendix A 'The Incompatibility of Fiducial and Bayesian Inference' to the text *The Estimation of Probabilities*, I.J. Good presents a paradox for fiducial inference based on data of the sort supposed in our case three.[22] Without going into the details of his counterexample, the distributions he considers are not capable of generating a smoothly-invertible, canonical pivotal; hence, do not represent a threat to fiducial probability as reconstructed herein. However, the strategy employed can be modified to a distribution with a 'natural' pivotal, so that the essence of the paradox survives. The motif of the counterexample is, as in Lindley's paradox, the incompatibility of the sequential methods when the data are reversed. That is, even without considering a third fiducial probability, based on a single sufficient statistic, the other two may still be inconsistent. Lindley's results presuppose the existence of a combined sufficient statistic, so no appeal to his coherence conditions will help.

The problematic example is very easily described.[23] Suppose an experimenter is faced with a hollow cubic structure of unknown dimensions and is asked to determine a probability distribution for the volume of the cube. Two experimental procedures are available and the question is whether the data from the two can be combined into a single fiducial posterior probability for the volume, represented by the parameter v. The alternative, independent

procedures are as follows. A liquid of known density, say 1 unit weight/unit volume, is poured into the cube until the inside is filled up, and then the liquid is removed and weighed on a scale whose readings are known to be $N(\theta, 1)$; where θ is the true weight of the object so measured. The second procedure is to take a rigid rod of known density, say 1 unit weight/unit length, and lay it off against one of the edges of the cube. Cut the rod so that its length coincides with the length of the edge, and weigh the cut rod on the same scale (described above).

Readings from the first experiment will be normally distributed about ν, with a variance of 1. Readings from the second experiment will be normally distributed about $\nu^{1/3}$ also with unit variance. This latter result follows simply from the elementary fact that length^3 = volume.

There is no sufficient statistic for ν, based on both observations, as demonstrated below. However, each result, taken on its own, is suitable for a fiducial argument about ν. The required pivotal are:

$$v_1 = (w_L - \nu); \text{ where } w_L \text{ is the liquid's weight,}$$
$$v_2 = (w_R - \nu^{1/3}); \text{ where } w_R \text{ is the rod's weight.}$$

Beginning with either w_L or w_R, v_1 or v_2 (respectively) can be used to induce a fiducial 'prior' distribution for the other observation. Then, using Bayes' method, i.e. CC and Bayes' theorem, the other result can be incorporated to produce a posterior fiducial probability for ν based on both results. Unfortunately, the two methods do not generate the same posterior probability. That this is so can be readily seen from the corresponding Bayesian analysis, due to Lindley. The conjugate prior for each of the two fiducial distributions generated by the above pivotals is the improper, uniform distribution for the parameter that appears in the pivotal. That is, for v_1 the conjugate prior is $d\nu$, while for v_2 it is $d\nu^{1/3}$. These are, of course, different and incompatible prior probabilities for the same unknown, ν. Hence, the sequential method *is not* invariant to the order of the conditioning argument. But, as the reader knows, Bayes' theorem *is* invariant to the order of conditioning.

The resolution I propose to this dilemma is, admittedly, inadequate in a serious respect. Instead of explaining away the paradoxical air of the counterexample and saving the reconstructed fiducial argument, the blame for the problem is located with the inherent sensitivity of the Principle of Insufficient Reason to even minor alterations in the formulation of the unknown parameter. That is, Fisher's admonition is to be recognized and supported if the fiducial argument is to be defended: but to do so is to restate the need for

extreme caution when applying fiducialized version of the old Laplacean principle.

Assessing the fiducial argument from the perspective of the Bayesian analysis, one finds that the various constraints on the pivotal, summarized as 'naturalness' of the pivotal, serve to pick out the conditions for application of an 'ignorance' prior. This conjugate prior probability is, by Lindley's results, known to be the improper, uniform distribution – which looks exactly like the prior recommended by the Principle of Insufficient Reason. In other words, the 'ignorance' is distributed, a priori, equally among all the epistemic possibilities, subject to the conditions of the fiducial argument.

Criticism of the Indifference Principle is very old, and very persuasive. In the spirit of some of the remarks of Boole and Venn on this point, a typical difficulty is illustrated in the following example; called (for lack of a better name) the 'wine-water paradox.'

Suppose that a liter of a wine-water mixture is known only to contain between 50% and 90% wine. Therefore, the ratio of wine to water varies from 1:1 to a maximum of 9:1. Applying the Principle of Insufficient Reason, the prior probability for the contents of the mixture is uniform over the various possibilities, which form a continuum of values. However, expressed in terms of percentage wine, the prior probability is $\frac{1}{2}$ that there is more than 70% wine; whereas, expressed in terms of the ratio of wine to water, the prior probability is $\frac{1}{2}$ that the ratio is greater than 5:1. But this last value corresponds to a minimum of at least (approximately) 83.5% wine. Unrestrained application of the Principle of Indifference leads to inconsistent results.

As is well known to those who attempt a modified use of this principle, little difficulty is experienced if the unknown parameter (for continuous distributions) can be limited to a canonical set of alternative forms. Practically, a family of linear transformations often is admissible without fear of inconsistency. Thus, for example, the switch from the wine-water ratio to the inverse, water-wine ratio would be barred, as it employs a non-linear transformation ($T: x \rightarrow x^{-1}$), with corresponding inconsistent 'uniform' prior probabilities.

Reconsider the paradox for the fiducial argument using the two sets of data. As pointed out on the previous page, the parameterization corresponding to these different data obtained from the fiducial argument are ν and $\nu^{1/3}$. These are not related by any linear transformation, and one suspects that the underlying problem is the same as that in the wine-water paradox, above. This analysis can be defended if Fisher's condition, i.e. the existence of a common sufficient statistic for the two data, is activated.

In place of a one parameter problem with representation v and $v^{1/3}$, suppose (for simplicity) that the reading are of two *different* cubic structures, with volumes v and v'. This converts the problem to bivariate normal, centered about the two observed readings with the two location parameters v and $v'^{1/3}$ (say). (See p. 151, for details of this solution.)

Imagine that the data for this problem are augmented by information connecting the two volumes, e.g. that $v = f(v')$; where f is some function. For instance, it might be learned that the two structures are of equal (unknown) volume, i.e. that f is the identity function. This converts the problem from bivariate to univariate status.

Now, some functions f produce mathematical representations which admit a common sufficient statistic for the parameter. In fact, for the normal distribution above, if the function is linear, i.e. the relation between v and v' is graphed as a straight line, then there is a sufficient statistic of the joint data; reducing the problem to a previous case and allowing Lindley's coherence results to be applied. For instance, if f is the identity map, then the average of the two data is a common sufficient statistic. (This is nothing more than the restatement that the sample average is a sufficient statistic of the full sample data, only using an n-dimensional representation in place of the two-dimensional case, described above.)

Thus, Fisher's assumption of a common sufficient statistic is guaranteed if the parameters are linearly related.[24] Hence, the use of a common sufficient statistic limits the alternative formulations for the parameter under investigation; so that the selection of an 'ignorance' prior can be made without fear of violating the probability calculus, as in the wine-water problem.[25]

The foregoing discussion, although supporting Fisher's intuition that there must exist a sufficient (or exhaustive) statistic in order that the fiducial argument be consistent, reveals the severely diminished plausibility of the whole program. Unless it is known, from the start, that the incoming data admit of suitable summarization (as a whole), the use of the fiducial argument on any initial part opens the way for un-Bayesian fiducial consequences, as demonstrated by the 'paradox.' That is, if the experimenter observes w_L in the morning, and uses it to induce a fiducial probability for the volume, then when he observes w_R he is committed (by the probability calculus and CC) to a posterior probability for v which differs from the one he would have held had the experiments been reversed.

With this exception, our reconstruction of the one parameter fiducial argument is to be accepted for the purpose of additional inquiry into the multiparameter fiducial argument. This is done for the sake of several

important epistemological considerations, which are independent of the sole weakness in the reconstructed, one parameter fiducial argument discussed above.

NOTES

[1] R.A. Fisher, 'Inverse Probability', *Proc. Cambridge Phil. Soc.* 26 (1930), 528–35. Rep. in *CMS*, paper #22.

[2] The fault does not rest with Fisher. He recognized a difference very early on: Fisher, Comments on Neyman's paper 'On the Two Different Aspects of the Representative Method', in *SESPJN*, pp. 136–8. However, papers such as: C.J. Clopper and E.S. Pearson, 'The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial', *Biometrika.* 26 (1934), 404–13, did serve to confuse the issues.

[3] L. Savage, 'Discussion of Conference on Fiducial Probability', *Bull. Inst. Int. Stat.* 40^2 (1963), 925.

[4] I. Hacking, *Logic of Statistical Inference* (Cambridge: Cambridge University Press, 1965), p. 137.

[5] See R.A. Fisher, *Smoking The Cancer Controversy*, (London: Oliver and Boyd, 1959). In the text I have used 'recognizability' to include both (i) and (ii) above. This differs from Fisher's later usage, when he distinguished 'recognizable' from 'relevant.' Therefore, read 'recognizable' as 'recognizable and relevant.'

[6] Compare this to the situation with Reichenbach's 'narrowest' reference class rule, discussed on p. 31.

[7] See the example of p. 51.

[8] Fisher, so far as I know, nowhere states these conditions explicitly; however, his remarks in *Statistical Methods and Scientific Inference*, pp. 73–74, 179, support the textual reading.

[9] John W. Tukey, 'Some Examples with Fiducial Relevance', *Annals of Math. Stat.* 28 (1957), 687–95. See Appendix 9.2 for proof of the equivalence of 'smooth invertibility' with Fisher's (i) unboundedness and (ii) monotonicity requirements. For discrete distributions condition *C* (above) is dropped.

[10] Fisher, *op. cit.*, pp. 169–75.

[11] *Ibid.*, pp. 127–32. Also, Fisher, 'Some Examples of Bayes' Method of the Experimental Determination of Probabilities A Priori', *J. Royal Stat. Soc., Ser. B* 24 (1962), 118–24.

[12] A.W.F. Edwards, *Likelihood* (Cambridge: Cambridge University Press, 1972), pp. 206–10.

[13] Allan Birnbaum, 'Concepts of Statistical Evidence,' in S. Morgenbesser, P. Suppes and M. White eds., *Philosophy, Science and Method* (New York: St. Martin's Press, 1969), pp. 112–143.

[14] The results for the case $t(x, y)$ equals *opposite* are also identical to the former, simultaneous ones.

[15] Fisher, 'On Some Extensions of Bayesian Inference Proposed by Mr. Lindley', *J. Royal Stat. Soc., Ser. B* 22 (1960), 299–301.

[16] D.V. Lindley, 'Review of Statistical Methods and Scientific Inference', *Heredity* 11 (1957), 280–3.

[17] Fisher, *op. cit.*, p. 301.
[18] D.V. Lindley, 'Fiducial distributions and Bayes' Theorem', *J. Royal Stat. Soc., Ser. B* 20 (1958), 102–7.
[19] Fisher, *op. cit.*, p. 299.
[20] Fisher, *Statistical Methods and Scientific Inference*, p. 58.
[21] Lindley, *op. cit.*, p. 299.
[22] I.J. Good, *The Estimation of Probabilities* (Cambridge: The M.I.T. Press, 1965), pp. 81–3.
[23] It has come to my attention that this problem was anticipated, in part, by L. Savage. See his remarks in The Discussion of Conference on Fiducial Probability, *Bull. Inst. Int. Stat.* 40² (1963), p. 927, about lines 3–9.
[24] Fisher has shown that, by use of the ancillarity principle to obtain exhaustive statistics, the functional relation between the parametric forms need not be linear. They might be, for example, circular. [Fisher, *Statistical Methods and Scientific Inference*, pp. 140–141.]
[25] One cannot employ the sequential method for including the general information that $\nu = f(\nu')$, for arbitrary functions f, analogous to the treatment of the data d: that $\theta \leqslant 15$, in the problem of p. 116. This is because, unlike that problem, once the fiducial argument is used to generate a joint density for the (ν, ν') parameter space, the conditional density for $\nu = f(\nu')$ is *undefined.* [This feature of the problem is similar to the well-known Borel paradox for conditional distributions on a sphere. See A.N. Kolmogorov, *Foundations of the Theory of Probability*, 2nd ed., trans, and ed. N. Morrison (New York: Chelsea, 1956), pp. 50–51.]

CHAPTER 5

THE FIDUCIAL ARGUMENT: SEVERAL PARAMETERS

Coming after the discoveries of Chapter 4, the problems discussed in Chapter 5 must appear anticlimatic and forbiddingly technical. We already know where the fatal flaw is in fiducial theory. Still, the picture remains incomplete without connecting the findings for the one-parameter case to the mass of published criticism involving multi-parametric fiducial inference. Some of the results, e.g. the discrete version of the Behrens-Fisher problem (5.1–5.2), are of technical interest primarily. Others, such as the connection between fudicial inference and recent 'marginalization' paradoxes (due to Dawid, Stone and Zidek), show up theoretically important questions raised by multivariate fiducial theory, questions which are still unanswered.

The opening section rehearses a 'step-by-step' method for generalizing the one-parameter fiducial argument, from Chapter 4. The generalization relies heavily on conditionalization, in a way not necessary for the one-parameter case. With one unknown factor the problem of summarizing all the evidence in a single sufficient (or exhaustive) statistic can be solved (if at all) noncontroversially. That is, with one parameter, the conditions under which a satisfactory summary of the data is possible avoid explicit reference to prior probability. A perfectly satisfactory frequentist definition is available. However, in the step-by-step method where parameters are dealt with *seriatim*, one must be able to match up statistics with parameters (for inference to marginal probability). No solution to this more vexing problem can be given without bringing prior probability into the analysis, and, of course, *conditionalization* is the approved method for carrying this out.

Much of the confusion that exists regarding multi-parametric fiducial inference is traceable directly to Fisher. Although, I claim, the step-by-step method is the only generally correct mode of fiducial inference with several unknowns (a claim with historical justification based on Fisher's last work), Fisher suggested alternative short-cut methods, e.g. a *direct* fiducial inference to 'Student's' t-distribution. An example, using a discrete distribution with the step-by-step method, illustrates how the *direct* method may not be available in all cases, (5.1). More importantly, this same example serves to clarify the independence between fiducial and confidence interval theories. The latter requires a repeated sampling interpretation, the former does not. As in the Behrens-Fisher problem, the fiducial solution for the discrete distribution problem fails to support a repeated sampling interpretation, (5.2). The generalization of multi-parametric fiducial inference concludes with a brief discussion of different distributional forms, and an application to the bivariate normal model, (5.3).

The reconstruction of the preceding three sections, 5.1–5.3, provides sufficient groundwork to enter into a detailed evaluation of published criticism, in the form of 'paradoxes,' surrounding fiducial inference. These are broken down into two major categories: conflicts with N-P theory (5.4); conflicts with Bayesian theory (5.5).

I have already argued that fiducial theory and even significance testing do not depend

upon a repeated sampling interpretation for their validity (3.1, 3.6, 4.3 and 5.2). Hence, the sting has been removed from objections to fiducial inference which make the mistake of infusing in it conflicting (and irrelevant) principles (5.4).

The more serious challenges, however, come from arguments purporting to show incoherence (usually through mutually conflicting patterns) of fiducial inferece. In response, I show that some combinations of the following considerations serve to explain away the difficulties: the step-by-step method may fail due to the lack of suitable statistics (for marginal inference); illigitimate transformations (3.4) may be the source of alternate, conflicting fiducial arguments; and (finally) an unwarranted use of a *conglomerability* principle (see also 9.1) may underline the 'paradox.' The upshot is that no new difficulties (beyond the one uncovered in 4.8) undermine fiducial inference. What problems remain, e.g. proper use and interpretation of 'marginal' probabilities with improper prior distributions, apply across the board. They challenge all theories of inference which incorporate 'ignorance' distributions (5.5).

5.1. MULTIVARIATE FIDUCIAL PROBABILITY AND THE STEP-BY-STEP PROCEDURE

With the exception of Lindley's work, discussed previously, almost all of the published analysis of the fiducial argument focuses on multivariate distributions and the related 'marginal' distributions for each parameter separately. Historically, the first clear example of the difference between fiducial and confidence methods (stemming from the 'repeated sampling' interpretation conflict) was the Behrens-Fisher problem, a bivariate normal question. Since 1954, with the appearance of the Creasy-Fieller paradox, most of the discussion has centered on inconsistencies in fiducial inference based on incompatible densities generated by alternative fiducial methods, e.g. sequential versus simultaneous techniques (see p. 120). The position to be presented here is that by carefully using the detailed conditions from the single parameter examples, and by precise use of CC, a coherent multivariate argument is available. In addition, the so called 'step-by-step' procedure, Fisher's last resource for the many paradoxes he faced, will be clarified and defended.

As in the previous chapter, the presentation will include both Fisher's analysis, and where that is found to be lacking, suggestions for making up for the deficiencies. The analysis concludes with a review of the more well known 'paradoxes.'

The final section of the text *Statistical Methods and Scientific Inference* contains a complex (five parameter) bivariate normal problem; an illustration of the fiducial argument for multiparameter distributions. Fisher ends the discussion with the remarks,

The example has been chosen to illustrate the process of building up the simultaneous distribution of the parameters rigorously by a step by step process, since the short cut suggested by the use of Jacobians has no claim to validity unless it can be proved to be equivalent to a genuine fiducial argument.[1]

This heuristic example has been the center of several disputes. The criticism, common to all, is that the 'step-by-step' method advocated by Fisher is susceptible to contradictions either from conflict within, i.e. by reversing the steps, or else from conflict with alternative 'direct' fiducial methods, i.e. by using different pivotals. Instead of jumping into the controversy directly, the strategy followed here is first to develop the 'step-by-step' technique for the simple two-parameter case, building on the reconstructed one-parameter case, in the hope that the understanding thus engendered will permit satisfactory resolution of the more complex difficulties.

If $\mathbf{p}(x, y/\theta_1, \theta_2)$ is the joint density function for the two random variables x and y, given the (unknown) parameters θ_1 and θ_2, then the multiplication theorems states that:

$$[\mathrm{I}] \qquad \mathbf{p}(x, y/\theta_1, \theta_2) = \mathbf{p}(x/\theta_1, \theta_2, y) \cdot \mathbf{p}(y/\theta_1, \theta_2),$$

and that

$$[\mathrm{II}] \qquad \mathbf{p}(x, y/\theta_1, \theta_2) = \mathbf{p}(x/\theta_1, \theta_2) \cdot \mathbf{p}(y/\theta_1, \theta_2, x).$$

Suppose that, given the background knowledge BK, the density function for the two random variables can be given as (suppressing BK):

$$[\text{Type } A] \quad \mathbf{p}(x, y/\theta_1, \theta_2) = \mathbf{p}(x/\theta_1, \theta_2) \cdot \mathbf{p}(y/\theta_2)$$

That is, using BK, the factorization I (from the multiplication theorem) can be modified. Specifically, BK entails that x and y are distributed independently, and that y depends on θ_2 only. The 'step-by-step' process for a joint fiducial probability, $\mathbf{p}(\theta_1, \theta_2/x, y)$, using a factorization of Type A, follows these three steps:

Step$_1$: Note, first, that given θ_2 x becomes sufficient for θ_1. That is, given θ_2, the distribution for x depends on θ_1 alone, so that (x, y) can be reduced to x without loss of relevant information (about θ_1). Suppose that, given θ_2, there exists a smoothly invertible, canonical pivotal between x and θ_1. In other words, given knowledge of θ_2, which reduces the problem to uni-variate status, there is a legitimate fiducial argument from x to θ_1. (Step$_1$ is very much like an assumption for proof by mathematical induction.)

Step$_2$: Extend the fiducial argument from x to θ_1 to cover cases where there exists a probability (any *proper* one will do) for possible values of θ_2

(now a nuisance parameter for θ_1) using the following conditionalizing argument. Since, for each value of θ_2 a fiducial density for θ_1, given x, exists, use the already existing probability density for θ_2 to weight the fiducial density for θ_1, now given $x \,\&\, \theta_2$. That is, Bayes' theorem states that

$$[\text{CF}] \quad \mathrm{p}(\theta_1/x) = \int_{\theta_2} \mathrm{p}(\theta_1/x \,\&\, \theta_2) \cdot \mathrm{p}(\theta_2) \cdot d\theta_2 .$$

Step_1 guarantees the existence of a legitimate fiducial argument based on the first factor to the right of the integral sign in CF (the conditional fiducial argument). The assumption of a probability for θ_2 (which usually is independent of x for factorization of Type A; hence x is ignored for updating $\mathrm{p}(\theta_2/BK)$) allows the use of CF to generate a 'marginal' (fiducial) density for θ_1. The joint density for $\Theta = (\theta_1, \theta_2)$ is, of course, just

$$\mathrm{p}(\Theta/x) = \mathrm{p}(\theta_1/x \,\&\, \theta_2) \cdot \mathrm{p}(\theta_2).$$

The epistemological underpinings for Step_2 are the CC principle, Bayes' theorem, and the requirement (for step_1) that θ_2 is known. But now in step_2, beliefs about θ_2 are merely probabilistic and not claims to certain knowledge. As in the one parameter cases, CC is the lynch pin (along with Bayes' theorem) for holding together the total set of beliefs, represented by the probability functions for the various unknowns.[2]

Step$_3$: Suppose that no independently available probability for θ_2 is available. That is, assume that the background knowledge is still weaker in that beliefs about θ_2 cannot be represented probabilistically, i.e. one is ignorant about both θ_1 and θ_2. The full, two-dimensional fiducial probability for $\Theta = (\theta_1, \theta_2)$ is completed by using $\mathrm{p}(y/\theta_2)$ to obtain a fiducial probability for θ_2 given y, *in the absence of knowledge of* θ_1.

Suppose that a smoothly invertible, proper pivotal for θ_2, given y, exists. The difficult part of the argument is demonstrating that, in the absence of knowledge of θ_1, y is sufficient for θ_2. This general problem has occupied the frequentists for many, many years. Dating as far back as 1935[3] and again recently,[4] this problem has surfaced as a central one for multiparameter inference. Most importantly, unlike in the simple one parameter case, the epistemic state for θ_1 (the nuisance parameter) is pertinent for the claim that y is sufficient for θ_2.[5] The upshot is that no clear counterpart to the one-dimensional sufficiency criterion (p. 112) is available.

The step-by-step technique does, however, allow for a coherence check

of this critical claim, essential for $step_3$. Specifically, $step_1$ commits the Bayesian analogue of the fiducial argument to a conjugate prior (the 'ignorance' prior) for θ_1. This prior can be used to test whether, in fact, y is sufficient for θ_2 in the absence of knowledge of θ_1. In other words, presupposing the joint fiducial density for $p(\theta_1, \theta_2/x, y)$, the 'marginal' density for θ_2 is obtained by 'integrating out' the nuisance parameter θ_1 (see p. 143 and also 158–9). The sufficiency claim is granted, i.e. the test proves coherence, if the resulting 'marginal' density is just the fiducial density obtained from the $step_3$ pivotal using $p(y/\theta_2)$. The unusual aspect of this $step_3$ fiducial argument is that the pivotal looks like an ordinary one parameter pivotal, i.e. being a function of y and θ_2 only. Nonetheless, θ_1 enters the analysis through the coherence check of the sufficiency claim. As will be demonstrated (p. 160), this check eliminates some (but not all) of the Dawid, Stone and Zidek paradoxes from the joint fiducial probability arguments.

Unfortunately, the Objective Bayesians cannot use this solution. To preview how their position differs, it is to be noted that the Objective Bayesians, e.g. Jeffreys, legislate in favor of improper 'ignorance' priors by way of some modified version of the Lapalcean Principle of Indifference. The modifications, needed to avoid paradoxes on the order of the 'wine-water' example (p. 12), take into consideration symmetries or subjective indifferences for the unknown parameter(s). These factors do not segregate out the multiparameter distributions (using improper prior densities) which fail the coherence check from those that pass it.[6] On the other hand, the step-by-step fiducial argument, which extracts warranted 'ignorance' prior only through frequency based considerations, *is* sensitive to this problem.

Granting, then, the sufficiency claim (backed up by the coherence check), we replace $p(\theta_2) \cdot d\theta_2$ from the conditional fiducial argument [CF] of $step_2$ by a fiducial probability generated from $p(y/\theta_2)$. The last term in CF has been eliminated by the fiducial argument, completing the bivariate fiducial probability for Θ.

As in previous sections, the argument is illustrated by a double example. The continuous one, for the simple bivariate normal $N(\mu, \sigma^2)$, is taken from Fisher's text *Statistical Methods and Scientific Inference*.[7] The discrete example, although not very elegant, is a continuation of Hacking's binomial problem. My excuse for it is that it permits a handy demonstration of properties of a marginal fiducial probability which duplicate properties of the Behrens-Fisher distribution.

Example (discrete): *Background Knowledge* – Among other items, it is

assumed that a coin is known to be binomial with parametric (bias) value θ equaling .2 or .4 or .6 or .8. Observations consist of two parts (x, y). The x ingredient is the 'heads' $[H]$ or 'tails' $[T]$ outcome of a binomial trial, with $\mathrm{p}(H/\theta = c) = c$. The y component is the observation of the denomination of the coin, either 'dime' $[d]$, or 'nickel $[n]$. No frequency based prior probability for θ is available; hence, the fiducial argument is in order.

Methodology: to keep the sample and parameter spaces of the same dimension, helping out with the smooth invertibility condition (B), i.e. 1-1 ness, transform the parameter space as follows:

Let the bias [B] of the coin be

$$\begin{pmatrix} + \text{ iff } \theta > .5, \text{ i.e. } \theta = .6 \text{ or } .8 \\ - \text{ iff } \theta < .5, \text{ i.e. } \theta = .4 \text{ or } .2 \end{pmatrix}$$

Let the loading [L] be

$$\begin{pmatrix} h \text{ (heavy) iff } \theta = .2 \text{ or } .8 \\ l \text{ (light) iff } \theta = .4 \text{ or } .6 \end{pmatrix}$$

More Background Knowledge: It is also known that the stamping machine which makes the coins has a screw loose. It produces coins with an off-center of gravity. The farther from the center of gravity the heavier the loading of the coin. The machine has two outputs: heavy and light loaded coins. Because nickels are thicker than dimes, it requires a larger machine error to produce a heavily loaded nickel than a heavily loaded dime. All this can be summarized by the rich statistical knowledge that:

$$\mathrm{p}(n/\mathrm{L} = l) = \mathrm{p}(d/\mathrm{L} = h) = .9 \text{ and}$$

$$\mathrm{p}(n/\mathrm{L} = h) = \mathrm{p}(d/\mathrm{L} = l) = .1.$$

Therefore, $\mathrm{p}(x, y/\mathrm{B}, \mathrm{L}) = \mathrm{p}(x/\mathrm{B}, \mathrm{L}) \cdot \mathrm{p}(y/\mathrm{L})$; an instance of factorization of Type *A*.

To permit the simultaneous, parallel presentation of the two examples, consider, at this point, the background knowledge for the continuous case.

Example (continuous): *BK* — A set of n independent readings $(x_1, \ldots, x_n)$ are taken from the normal distribution $\mathrm{N}(\mu, \sigma^2)$. Both parameters (μ, σ^2) are unknown, and the background knowledge is insufficient for the determination of prior probabilities using frequentist principles.

Methodology: To keep the sample and parameter spaces of the same dimension, helping out with the smooth invertibility condition (B), i.e. 1-1

ness, transform the sample space of n-fold dimension to the two-dimensional $(\bar{x}, s)$, where $\bar{x}$ and s are defined as:

$$\bar{x} = \left(\sum_i x_i\right)\Big/ n; \quad \text{and}$$

$$s = \left(\sum_i [x_i - \bar{x}]^2/(n-1)\right)^{.5} \text{ (positive)}.$$

It is well known that $(\bar{x}, s)$ are jointly sufficient for (μ, σ^2). Also, $\mathrm{p}(\bar{x}, s/\mu, \sigma^2) = \mathrm{p}(\bar{x}/\mu, \sigma^2) \cdot \mathrm{p}(s/\sigma^2)$; which is an instance of factorization of Type A.

The Step-by-Step Method

Continuous

Step$_1$: Given σ^2, $\bar{x}$ is sufficient for μ. Let the pivotal v_μ be defined as: $v_\mu = (\bar{x} - \mu)/\sigma$. v_μ is distributed normally $\mathrm{N}(0, 1/n)$ independent of the unknown μ. v_μ is smoothly invertible, and canonical with $\mu^* = 0$. Therefore, there exists a fiducial argument from $\bar{x}$ to μ, given σ^2.

Step$_2$: Let $\mathrm{p}(\sigma^2) \cdot d\sigma^2$ be given so that CF: $\mathrm{p}(\mu/\bar{x}) = \int_\sigma \mathrm{p}(\mu/\bar{x}, \sigma^2) \cdot \mathrm{p}(\sigma^2) \cdot d\sigma^2$. If the probability for σ^2 is inversely proportional to a χ^2 distribution, then the 'marginal' fiducial probability for μ is 'Student's' t.

Discrete

Step$_1$: Given $\mathbf{L}$, x is sufficient for $\mathbf{B}$. Let the pivotal $v_{\mathbf{B}}$ be defined as:

(for $\mathbf{L} = l$) $v_{\mathbf{B}} =$

$$\begin{bmatrix} & \mathbf{B}=+ & \mathbf{B}=- \\ o_1 & \text{iff } (H \,\&\, \theta = .6) & \text{or } (T \,\&\, \theta = .4) \\ & \mathbf{B}=+ & \mathbf{B}=- \\ o_2 & \text{iff } (T \,\&\, \theta = .6) & \text{or } (H \,\&\, \theta = .4) \end{bmatrix}$$

(for $\mathbf{L} = h$) $v_{\mathbf{B}} =$

$$\begin{bmatrix} & \mathbf{B}=+ & \mathbf{B}=- \\ o_1 & \text{iff } (H \,\&\, \theta = .8) & \text{or } (T \,\&\, \theta = .2) \\ & \mathbf{B}=+ & \mathbf{B}=- \\ o_2 & \text{iff } (T \,\&\, \theta = .8) & \text{or } (H \,\&\, \theta = .2) \end{bmatrix}.$$

$v_{\mathbf{B}}$ is distributed binomially, $\theta = .6$ $(.8)$ for $\mathbf{L} = l$ (h). $v_{\mathbf{B}}$ is smoothly invertible, and canonical with $\theta^* = .6$ $(.8)$, when $\mathbf{L} = l$ (h), respectively. Therefore, there exists a fiducial argument from x to $\mathbf{B}$, given $\mathbf{L}$.

Step$_2$: Let $\mathrm{p}(\mathbf{L}/BK)$ be given, so that CF: $\mathrm{p}(\mathbf{B}/x) = \Sigma_{\mathbf{L}} \mathrm{p}(\mathbf{B}/x, \mathbf{L}) \cdot \mathrm{p}(\mathbf{L}/BK)$.

Continuous

Step$_3$: Assume that s is sufficient for σ^2, in the absence of μ. This claim can be checked for coherence by using the conjugate prior for μ from step$_1$, i.e. the improper, uniform $d\mu$, and through CC and Bayes' with the fiducial probability for σ^2, given s. Define the pivotal v_{σ^2} as: $v_{\sigma^2} = [n-1] \cdot [s^2/\sigma^2]$. Then v_{σ^2} is distributed as χ^2 with $(n-1)$ degrees of freedom, independently of the unknown variance. v_{σ^2} is smoothly invertible, and canonical with $\sigma^{2*} = 1$. Thus, a fiducial probability for σ^2, given s in the absence of knowledge of μ, is available. The joint distribution $\mathrm{p}(\mu, \sigma^2/\bar{x}, s)$ is obtained by ignoring the integral in CF, and using v_{σ^2} to induce a (fiducial) probability for σ^2. Since this distribution is inversely proportional to χ^2, the 'marginal' fiducial probability for μ is given by 'Student's' t, with $(n-1)$ degrees of freedom.

Discrete

Step$_3$: Assume that y is sufficient for **L**, in the absence of knowledge of **B**. This claim can be checked for coherence by using the conjugate prior for **B**, from step$_1$, i.e. $\mathrm{p}(+) = \mathrm{p}(-) = .5$, and through CC and Bayes' with the fiducial probability for **L**, given y. Define the pivotal $v_{\mathbf{L}}$ as:

$$v_{\mathbf{L}} = \begin{bmatrix} o_1' \text{ iff } (n \,\&\, \mathbf{L} = l) \text{ or } (d \,\&\, \mathbf{L} = h) \\ o_2' \text{ iff } (d \,\&\, \mathbf{L} = l) \text{ or } (n \,\&\, \mathbf{L} = h) \end{bmatrix}$$

$v_{\mathbf{L}}$ is binomially distributed with $\theta = .9$, independently of the unknown Load **L**. $v_{\mathbf{L}}$ is smoothly invertible, and canonical with $\mathbf{L}^* = l$ or $\mathbf{L}^* = h$, i.e. y is binomial with $\theta^* = .1$ or $\theta^* = .9$. Thus a fiducial probability for **L**. given y and in the absence of knowledge of **B**, is available. The joint distribution $\mathrm{p}(\mathbf{B}, \mathbf{L}/x, y)$ is obtained by ignoring the summation in CF, and using $v_{\mathbf{L}}$ to induce a (fiducial) probability for **L**.

The results for the discrete example are summarized in Table I, below. Reading vertically, one obtains the conditional probability $\mathrm{p}(x, y/\mathbf{B}, \mathbf{L})$. Reading horizontally, one obtains the posterior joint fiducial probability $\mathrm{p}(\mathbf{B}, \mathbf{L}/x, y)$, derived through the step-by-step procedure.

TABLE I

(B, L)	$+h$	$-h$	$+l$	$-l$
(x, y)				
Hn	.08	.02	.54	.36
Hd	.72	.18	.06	.04
Tn	.02	.08	.36	.54
Td	.18	.72	.04	.06

One can readily see that y is sufficient for **L** *only* in the absence of knowledge of **B**. If **B** were known, x would be relevant to the posterior probability for **L**. For example, $\mathrm{p}(l/H, n) = \mathrm{p}(l/n) = .9$ so long as **B** is unknown. However,

if $\mathbf{B} = +$ is known, then $\mathrm{p}(l/H, n, +) = .54/(.54 + .08) \neq \mathrm{p}(l/n, +) = (.62) \cdot (.54/[.54 + .08]) + (.38) \cdot (.36/[.36 + .02]) = .9$

The Bayesian model for this family of probabilities has conjugate priors, $\mathrm{p}(+) = \mathrm{p}(-) = \mathrm{p}(h) = \mathrm{p}(l) = .5$, with independence between **B** and **L**, i.e. $\mathrm{p}(\mathbf{B}, \mathbf{L}) = \mathrm{p}(\mathbf{B}) \cdot \mathrm{p}(\mathbf{L})$. Hence, y is sufficient for **L** just in case the conjugate prior for **B** is the uniform, ignorance prior $\mathrm{p}(+) = \mathrm{p}(-) = .5$.

As reported previously, one of the most common reasons for dissatisfaction with the multivariate fiducial argument is the fact that alternative methods yield incompatible joint fiducial probabilities. A principal source of these substitute techniques is Fisher's own writings. For example, with respect to the marginal fiducial probability for μ (from the continuous case) when σ^2 is unknown, he writes in *Statistical Methods and Scientific Inference*,

> In the test of significance due to 'Student' (W.S. Gossett), and generally known as the *t*-test, the data are taken to consist of N values of a single observable variate x, which are to be interpreted on the hypothesis that they are independent values of a variate normally distributed, with both mean and variance unknown, and of which no probability statements *a priori* are available. The two statistics $(\bar{x}, s^2)$ are known to be jointly Sufficient for the estimation of the true mean, μ, and the true variance, σ^2. Their sampling distributions in random samples of N may be expressed exactly in terms of the population parameters. . . .
>
> On the basis of these distributions 'Student' set himself to ascertain the exact sampling distribution of the ratio between the sampling error of the mean, and its standard error *as estimated*, namely
>
> $$t = [\sqrt{N} \cdot (\bar{x} - \mu)]/s \tag{35}$$
>
> when $\bar{x}$ and s are calculated from a finite sample, and he was successful in establishing the frequency element
>
> . . . . . (36)
>
> a distribution depending on N only, independent of both parameters of the Normal distribution sampled, and which can, therefore, be tabulated so as to give the value of t exceeded in absolute value, with any given probability, and for any number, $n = (N - 1)$, of the degrees of freedom.
>
> It will be recognized that 'Student's' distribution allows of induction of the fiducial type, for the inequality
>
> $$\mu < (\bar{x} - ts/\sqrt{N}), \tag{37}$$
>
> will be satisfied with just half the probability for which t is tabulated, if t is positive, and with the complement of this value if t is negative. The reference set for which this probability statement holds is that of the values of $\mu, \bar{x}$ and s corresponding to the same sample, for all samples of a given size or all normal populations. Since $\bar{x}$ and s are jointly Sufficient for estimation, and knowledge of μ and σ *a priori* is absent, there is no possibility of recognizing any subset of cases, within the general set, for which any

different value of the probability should hold. The unknown parameter μ has therefore a frequency distribution *a posteriori* defined by 'Student's' distribution.[8]

Fisher gives what appears to be a short-cut around the step-by-step method for generating the marginal fiducial distribution for μ. The pivoral t (line 35 from Fisher's statement) appears to be a one-parameter pivotal from the data $(\bar{x}, s)$ to parameter μ, which results from the real one-parameter pivotal (when σ is known) by substituting s for σ. One might even argue that t is a canonical pivotal for the statistic $\bar{x}/s$ when $\mu = \mu^* = 0$. It must be emphasized, however, that in the absence of knowledge of the variance σ^2, the problem is multiparametric and probabilities for the unknown mean μ are 'marginal.' Henceforth, label those attempted fiducial arguments, like the one using Student's t-variable, which induce 'marginal' fiducial densities without appeal to any step-by-step procedure. *'direct' fiducial arguments.*[9] The reconstruction given here is that 'direct' fiducial arguments depend for their validity on an agreement with a rigorous step-by-step fiducial argument. Moreover, I believe that Fisher saw the 'direct' argument as only provisionally correct: the condition being the availability of a supporting step-by-step alternative derivation.

Some forty pages after his first remarks on Student's distribution and its use for deriving fiducial probabilities about μ without knowing σ^2, Fisher reviews the subject of inference with a normal distribution (μ, σ^2) and adds the following:

> The rigorous step-by-step demonstration of the bivariate distribution by the fiducial argument would in fact consist of the establishment of the second factor giving the distribution of σ given S, disregarding the other parameter, μ, and then of finding the first factor as the distribution of μ given $\bar{x}$ and σ. Several writers have adduced instances in which, when the formal requirements of the fiducial argument are ignored, the results of the projection of frequency elements using artificially constructed pivotal quantities may be inconsistent. When the fiducial argument itself is applicable, there can be no such inconsistency.[10]

The argument is both clear and accurate. Apart from the revised order of 'steps' (which can be accounted for because the presentation given here follows an 'espistemological' chain of events, whereas Fisher's line facilitates calculations), there is no difference in method. Furthermore, the importance of relying on the step-by-step method is emphasized when mistakes, resulting from abuse of the 'direct' method, arise. The case in point concerns a faulty analysis, due to (of all people) D.V. Lindley. Fisher says, concerning the joint distribution obtained by the step-by-step process, that

> It will not be noticed that in this simultaneous distribution (83) μ and σ^2 are not

distributed independently. Integration with respect to either variable yields the unconditional distribution of the other, and these are naturally those obtainable by direct application of the fiducial argument, namely that

$$(\mu - \bar{x})/s \tag{84}$$

is distributed as is t for $(N-1)$ degrees of freedom, while

$$S/\sigma^2 \tag{85}$$

is distributed as is χ^2 for $(N-1)$ degrees of freedom. The distribution of any chosen function of μ and σ^2 can equally be obtained.[11]

However, Lindley, in his review of the first edition of *Statistical Methods and Scientific Inference*, neglects this underpining of the 'direct' fiducial method and is led to an unwarranted criticism of fiducial probability.[12] Believing that the t-pivotal is all that is relevant about μ when σ^2 is unknown, he points out, *correctly*, that a coherent probability for μ based on two separate samples and therefore two separate t values cannot be obtained by using the 'direct' fiducial method alone. That is, despite the apparent simplicity of the 'direct' fiducial argument through the t pivotal, two independent samples cannot be combined (properly) by considering the t values alone.

Fisher and Yates reply[13] (separately) that the correct fiducial 'marginal' probability for μ, based on both samples, is derived by the step-by-step method; which yields a coherent set of values. As expected, this corrected result agrees with a 'direct' application of the fiducial argument if both samples are combined first. The source of Lindley's error is found with the aid of the rigorous step-by-step method. As Fisher points out in the paragraph quoted last, the joint fiducial posterior probability for μ and σ^2 is not independent in the two parameters; hence, both must be used for constructing 'marginal' probabilities. Lindley's try at constructing a marginal probability for μ, based on two samples, neglects this technical fact. That it was overlooked is not surprising, given that Lindley considered only the 'direct' t-pivotal method suggested by Fisher.

5.2. MULTIVARIATE FIDUCIAL PROBABILITY AND REPEATED SAMPLING

A second and equally important problem concerning the 'direct' method and 'marginal' fiducial probability is the question of the existence of a 'repeated sampling' interpretation. Is it always the case that the 'marginal' fiducial probability can support a Neyman-Pearson 'repeated sampling' interpretation? The answer is no! Historically, the example which demonstrated this most

clearly was the Behrens-Fisher solution to the problem of the same name (p. 63).[14]

That the answer is no, can be made abundantly clear by reviewing the discrete bivariate example (p. 142). However, before proceeding with this part of the analysis, reconsider the connection between fiducial probability and significance testing. Marginal probabilities correspond to *composite* null hypotheses. For example, 'Student's' problem: the value of the unknown mean μ of a normal distribution in the absence of knowledge of the variance σ^2, treats composite null hypotheses of the form (c, σ^2); where c is the postulated value for μ, and σ^2 is left unspecified.

That 'Student's' problem admits of an 'exact' test of significance, i.e. one independent of the actual value of σ^2, shows why the corresponding ficucial 'marginal' probability allows a 'direct' method. The existence of an 'exact' significance test is, of course, insufficient to guarantee the existence of a 'direct' fiducial probability. For example, independence tests for 2 x 2 tables (p. 97) have exact tests but fiducial probability is not defined for these cases. However, if an exact significance test exists, then a 'direct' pivotal based on it can be used so long as the 'direct' method leads to results agreeing with CF fiducial probabilities, from step_2 of the step-by-step method (p. 139).

In 'Student's' problem, the exact significance test (t-test) for composite null hypotheses about the unknown mean supplies a 'direct' pivotal for fiducial inference about the mean. (See Fisher's comments quoted on p. 144.) A comparsion of the 'direct' fiducial probability with the marginal fiducial probability from the step-by-step method (pp. 142–3, the continuous case) shows the two solutions agree and the 'direct' method is validated. But, since the step-by-step method does not, in general, presuppose an exact significance test corresponding to CF probabilities there is no assurance that marginal fiducial probabilities support a 'repeated sampling' interpretation.

Now to the discrete bivariate example as illustration of this phenomenon. The posterior 'marginal' probability for **B** is given in Table II (below). Reading horizontally, one obtains the posterior fiducial probability for $\mathbf{B} = (+, -)$. The vertical numbers have no significance as 'sampling' or 'repeated sampling' probabilities. One feels the force of the situation by seeing that the posterior fiducial probability for $\mathbf{p}(+/Hn)$, which equals .62, does not equal the direct probability for any sampling statistic. That is, reading vertically in either column 1 or column 3 of Table I (p. 143), no function (reduction) of the observation has a sampling distribution of .62. Hence, there is no confidence interval analysis compatible with Table II.

TABLE II

B	+	−
(x, y)		
Hn	.62	.38
Hd	.78	.22
Tn	.38	.62
Td	.22	.78

In the step-by-step procedures discussed, **B** is analogous to μ, and **L** to σ^2. Both distributions are bivariate with factorizations of Type *A*. Nonetheless, only the continuous example produces a 'marginal' probability compatible with a 'repeated sampling' interpretation. Since the Neyman-Pearson interpretation can be used for one conclusion and not the other (and both are valid fiducial inferences), the 'step-by-step' procedure shows up yet another source of conflict between Fisherians and Neyman-Pearsonites.

Historically, this case displayed itself in a more complex distribution, the bivariate normal with correlation $\rho_{xy} = 0$ and the other four parameters, $\mu_{\bar{x}}$, $\mu_{\bar{y}}$, $\sigma^2_{\bar{x}}$, $\sigma^2_{\bar{y}}$ all unknown. The Behrens-Fisher problem, i.e. the question about the difference $\mu_{\bar{x}} - \mu_{\bar{y}}$, has a fiducial solution which cannot support a 'repeated sampling' interpretation. If $\delta = (\mu_{\bar{x}} - \mu_{\bar{y}})$ and $\zeta = (\sigma^2_{\bar{x}}/\sigma^2_{\bar{y}})$, then transformation to (δ, ζ) allows a factorization of Type *A* using Student's distribution and the jointly sufficient statistics (d', z), defined as follows. Let $d = \bar{x} - \bar{y}$, then

$$z = (s^2_{\bar{x}}/s^2_{\bar{y}}) \text{ and } d' = d \Big/ \left[\frac{s^2_{\bar{x}} + s^2_{\bar{y}} \cdot \zeta}{2} \cdot \frac{(\zeta + 1)}{-\zeta}\right]^{.5}.$$

Specifically: $\mathrm{p}(d', z/\delta, \zeta) = \mathrm{p}(d'/\delta, \zeta) \cdot \mathrm{p}(z/\zeta)$ where $\mathrm{p}(d'/\delta, \zeta)$ has a Student's distribution centered about δ, with $N_{\bar{x}} + N_{\bar{y}} - 2$ degrees of freedom, and $\mathrm{p}(z/\zeta)$ has Fisher's F distribution, independent of all parameters.

As in the discrete example, there is no 'direct' fiducial argument from (d', z) to hypotheses about δ, since there is no exact significance test for (composite) hypotheses about δ which is independent of the unknown ζ. The Behrens distribution (of the Behrens-Fisher solution) involves the statistic z, as well as d', for calculating the posterior, marginal fiducial probability for δ. For comparison, a Behrens-like distribution for the discrete example is available and is given below.

For variables (u, w):

$$\text{let } \mathrm{p}(u_1) = .62 \text{ and } \mathrm{p}(u_2) = .38 \quad \text{if } w = w_1 \text{ and}$$
$$\text{let } \mathrm{p}(u_1) = .78 \text{ and } \mathrm{p}(u_2) = .22 \quad \text{if } w = w_2.$$

Now, if the y observation is used to determine the w entry in the (u, w) distribution (analogous to z in the Behrens-distribution), then there is a natural pivotal between B and x which has the conditioned u distribution. It induces the marginal probability for B (of Table II, p. 148), given (x, y). Specifically, let pivotal v'_{B} be defined as:

$$v'_{\mathrm{B}} = \begin{pmatrix} o_1 \text{ iff } (+ \,\&\, H) \text{ or } (- \,\&\, T) \text{ when } w = w_1 = \text{nickel} = y \\ o_2 \text{ iff } (+ \,\&\, T) \text{ or } (- \,\&\, H) \text{ when } w = w_1 = \text{nickel} = y. \end{pmatrix}$$

$$v'_{\mathrm{B}} = \begin{pmatrix} o_1 \text{ iff } (+ \,\&\, H) \text{ or } (- \,\&\, T) \text{ when } w = w_2 = \text{dime} = y \\ o_2 \text{ iff } (+ \,\&\, T) \text{ or } (- \,\&\, H) \text{ when } w = w_2 = \text{dime} = y. \end{pmatrix}.$$

Then v'_{B} has all the desired properties for the Behrens-like distribution.

5.3. ADDITIONAL FACTORIZATIONS AND THE FIVE PARAMETER BIVARIATE NORMAL

The step-by-step procedure for generating bivariate fiducial arguments may be generalized to other distributions, besides factorizations of Type *A* (p. 138), by making use of the exhaustive step of the uni-parameter reconstruction.

[Type *B*] $\mathrm{p}(x, y/\theta_1, \theta_2) = \mathrm{p}(x/y, \theta_1, \theta_2) \cdot \mathrm{p}(y/\theta_2)$.

If the background knowledge contains enough information to permit factorization of Type *B* (compare this to I of p. 138), then the fiducial argument may be derived using an exhaustive statistic in step$_1$. With this factorization it follows that, given θ_2, x is exhaustive and y ancillary for θ_1; so that a conditional fiducial probability from x to θ_1, given θ_2, may exist. No new principle is needed for this extension.

[Type *C*] $\mathrm{p}(x, y/\theta_1, \theta_2) = \mathrm{p}(x/\theta_1, \theta_2) \cdot \mathrm{p}(y/x, \theta_2)$

If the background knowledge contains enough information to permit factorization of Type *C* (compare this to II of page 138), then the fiducial

argument may be derived using a *modified* exhaustive statistic in step$_3$. With this factorization it follows that in the absence of knowledge of θ_1, y is exhaustive, conditioned on x, for θ_2. Call such a statistic 'quasi-sufficient.' Obviously, since there is no frequentist criterion for determining the simpler claim, i.e. that y is sufficient for θ_2 (in the absence of knowledge of θ_1) with factorizations of Type A, no simple test for 'quasi-sufficiency' exists. As before, all that is available is a coherence check, analogous to the check at step$_3$ with factorization of Type A (p. 140), using the conjugate prior for θ_1 determined from the assumption at step$_1$.

[Type O] $\quad \mathbf{p}(x, y/\theta_1, \theta_2) = \mathbf{p}(x/\theta_1) \cdot \mathbf{p}(y/\theta_2)$

Factorizations of Type O involve only trivial multivariate fiducial inference. No additional machinery beyond the simple 'one-dimensional' case is needed. That is, no step-by-step process is necessary. Each factor to the right of the equality sign is assumed capable of generating a simple fiducial probability for the unknown parameter appearing in it. The joint density is merely the product of the two separate densities, and the parameters are distributed independently, a posteriori.

Factorization of Type O does not have much theoretical importance, but it does play a major role in the reconstruction of the bivariate normal distribution, a topic of considerable practical importance. Also, Fisher's concern with the step-by-step process focused on the bivariate normal distribution, as that is where many of the alleged paradoxes (to be discussed in the next section) were 'discovered.' Therefore, the step-by-step process for this distribution will be sketched, below.

The bivariate normal involves five parameters; two means μ_x and μ_y; two variances σ_x^2 and σ_y^2; and the correlation coefficient ρ_{xy}. There exists a set of five jointly sufficient statistics for the five parameters; two sample means $\bar{x}$ and $\bar{y}$; two sample variances s_x^2 and s_y^2 (with denominators $[n-1]$, not n); and the sample correlation coefficient $r = [\Sigma_i(x_i - \bar{x})(y_i - \bar{y})]/(n-1)s_x s_y$, $1 \leqslant i \leqslant n$ and data (x_i, y_i).

The parameters are grouped into a three levelled hierarchy.[15] First, the means are left unknown, while the variances and correlation are known. Then, weakening the background, the variances are considered unknown with (i) the ratios known, and (ii) the ratios unknown. Finally, the correlation coefficient is left unknown. Other possible epistemic situations, e.g. variances known, means unknown but known equal and correlation unknown, are quite complicated, perhaps even *without* fiducial counterparts, and will not be treated here.

If the correlation ρ_{xy} and variances σ_x^2 and σ_y^2 are known, a simple transformation yields a factorization of Type O for the unknown (transformed) parameters relating to the unknown means. Let:

$$w = \bar{x}/\sigma_{\bar{x}} + \bar{y}/\sigma_{\bar{y}} \quad \text{and} \quad z = \bar{x}/\sigma_{\bar{x}} - \bar{y}/\sigma_{\bar{y}}, \quad \text{and}$$

let: $$\gamma = \mu_x/\sigma_{\bar{x}} + \mu_y/\sigma_{\bar{y}} \quad \text{and} \quad \delta = \mu_x/\sigma_{\bar{x}} - \mu_x/\sigma_{\bar{x}}.$$

Then $p(w, z/\gamma, \delta)$ admits a factorization of Type O:

$$p(w, z/\gamma, \delta) = p(w/\gamma) \cdot p(z/\delta),$$

which also allows a simple fiducial argument for both unknown parameters, given the background knowledge specified above.[16]

Moving to the second level, with the variances unknown but variance ratio known, we find a simple step-by-step argument (analogous to the continuous example of p. 142) permits a 'Student's' distribution for the 'marginal' fiducial distributions (μ_x, μ_y), based on $(N_x + N_y - 2)$ degrees of freedom.[17]

If, in addition, the variance ratios are unknown, then a multivariate analogue of the Behrens-Fisher problem (with general correlation, not just $\rho_{xy} = 0$) is generated. An extra conditioning step is needed to get back the previous stage, with variance ratios known.

Finally, with all five parameters unknown, the fiducial argument begins with the derivation of the probability for ρ_{xy}, given r. This was the first fiducial argument, presented in 1930. Once this probability is determined, a CF probability for the variance ratio is available, bringing the analysis back to the previous level. This analysis corresponds to the treatment, given by Fisher, at the conclusion of *Statistical Methods and Scientific Inference* for the same distribution.[18]

The next, and final section in the reconstruction of the fiducial argument consists of a brief review of the more well known paradoxes. Instead of an individual case by case presentation, the paradoxes are grouped so that responses can be given generally, eliminating several with one swoop. Although the technical details of the preceding bivariate normal fiducial argument are not needed to follow the discussion, the reader is urged to refamiliarize himself with the step-by-step procedure and the definition of a canonical pivotal. [See pp. 138 and 125.]

5.4. REVIEW OF PARADOXES: (1) CONFLICTS WITH N-P THEORY

The criticisms of the fiducial argument can be grouped under two basic headings: conflict with Neyman-Pearson confidence intervals and conflict

with the Bayesian program. Those paradoxes which focus on the first problem will be considered in advance of the second, which tend to be more complicated.

The position adopted here is that when fiducial probability conflicts with Neyman-Pearson confidence theory, it is a result of the inability of the confidence theory to make proper use of all the data in-hand. The scope of the clash ranges from cases where (*a*) fiducial probability supports a 'repeated sampling' interpretation, but fairs only poorly with respect to the various operating characteristics used to evaluate the intervals, i.e. 'shortest' or 'unbiased' confidence intervals; (*b*) to cases where the fiducial argument does not even allow a 'repeated sampling' interpretation, e.g. the Behrens-Fisher problem. Examine conflicts of type (*a*) first.

That the confidence program is incapable of making full use of the data in-hand has been argued previously (see p. 56). That the two probabilities, i.e. fiducial and confidence, coincide for many cases is not surprising since, as already demonstrated, the 'best' tests, analyzed by operating characteristics, make use of sufficient statistics when they exist. Of course, fiducial probability also presupposes a sufficient, or at least exhaustive, summary of the data. Hence, sufficiency does not constitute a source of conflict.

The same is not true of exhaustive estimation, which relies on the ancillary statistic for conditioning. That ancillarity leads to a clash between fiducial and confidence analysis has been discussed on p. 118. A well-known case, exemplifying the incompatibility due to ancillarity, is B.L. Welch's 1939 example.[19]

Let (x_1, x_2) be two independently distributed variables from the uniform distribution $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, θ unbounded. The statistic $z_1 = (x_1 - x_2)/2$ has a distribution independent of θ, and is ancillary. Therefore, $z_2 = (x_1 + x_2)/2$ is conditionally sufficient (exhaustive) for θ, since (z_1, z_2) are logically equivalent data to (x_1, x_2). In other words, the transformation from (x_1, x_2) to (z_1, z_2) is 1-1 and calculable. A smoothly invertible, canonical pivotal for θ exists based on the exhaustive statistic z_2. It induces a fiducial distribution for θ such that the 'central' $(1 - \alpha)$ interval is defined as:

$$z_2 \pm (1 - \alpha) \cdot (.5 - |z_1|).$$

This interval carries a corresponding significance interval interpretation, where the significance test uses the ancillary statistic z_1, analogous to the fixed margins of the 2×2 test. That is, z_1 varies from trial to trial but the significance level is calculated with z_1 fixed. As in the bounded uniform

distribution case of p. 116, the surviving fiducial interval uses $100 \cdot (1-\alpha)\%$ of the possible θ values. Also, the example falls under category two, for the uni-variate case (p. 120), i.e. the data are each of the same kind and can be combined into a single exhaustive summary. The fiducial argument based on either one of the variables uses the smoothly invertible, canonical pivotal defined as: $v=(\theta-x)$, which is uniformly distributed on $[-.5, .5]$. Using Lindley's results, we see that the two fiducial methods are compatible.

Return to Welch's example. He has shown that the confidence analysis for the same problem does not make full use of the ancillary statistic, so that the confidence interval (at the $(1-\alpha)\%$ level) is given by:

$$z_2 \pm \text{minimum } \{(.5-|z_1|),(.5+|z_1|-[.5-(1-\alpha)/2]^{.5})\}.$$

As in the example of p. 56, this interval sometimes gives the entire possible range for θ (when $[1-2|z_1|]$ is small) at less than the 100% level. Nonetheless, the confidence interval is a 'best unbiased' one. The operating characteristics, specifically 'power,' fail to be sensitive to all the data available. The fiducial intervals, although supporting a 'repeated sampling' interpretation, do not get picked out as 'best' by the confidence interval theory.

Ancillarity is the source of yet another class of conflicting examples. One of the more controversial Neyman-Pearson methods, even for orthodox statisticians, is the 'mixed' or 'randomized' test (discussed on p. 42). As noted, randomization serves two basic purposes. It permits the 'smoothing out' of discrete distributions, hence allowing tests of all levels of significance. Also, it permits the combination of two different tests, of different size and power, into one test usually of better resolution than either of the other two taken alone.

The fiducial argument uses ancillarity in just the reverse process. That is, the conditioning step for exhaustiveness separates off cases, based on irrelevance. (See, for example, the remarks on p. 118.) The use of a randomizer to complete the 'mixed' test violates the ancillarity principle. That is, the outcome of the randomizer is irrelevant to the unknown parameter, since its distribution is *independent* (from the definition of the randomizer) of the parameter.

Fisher's predictable reaction to randomization in tests, as promoted by Neyman-Pearson techniques, is typified by his remarks on a mixed test solution for the Behrens-Fisher problem (p. 63),

> This proposal, which has perhaps now been abandoned (though at the time an equally faulty proposal was quickly put forward by Neyman) has two conspicuous objections, one of which is of general importance in that it applies to all 'randomization tests' in

the Natural Sciences. Namely, that if T and T' lie on opposite sides of the criterion, T' being always the larger, and a coin is thrown to decide which shall be chosen, it is then obvious at the time that the judgement of significance has been decided not by the evidence of the sample, but by the throw of the coin. It is not obvious how the research worker is to be made to forget this circumstance; and it is certain that he ought not to forget it, if he is concerned to assess the weight only of objective observational facts against the hypothesis in question.[20]

A good source for type (*b*) conflicts, where fiducial and confidence methods diverge, is the multiparameter situation. Looking at those questions which focus on one parameter (marginal probabilities) or transformations of the two parameters into a single composite parameter, we have seen (p. 147) that fiducial probabilities support a 'repeated sampling' confidence interpretation only if a 'direct' fiducial method is available. With this result in mind, reconsider the Master Significance Test for the bivariate normal distribution.

$$[\text{MST}] \quad \frac{a(\mu_x - x) + b(\mu_y - y)}{[a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho_{xy}]^{.5}}$$

From the well known properties of the bivariate normal distribution, MST is standard normal N(0, 1), for all a, b. By fixing values for a and b, we specify a family of homogeneous linear hypotheses about the unknown means. For example, using MST, we can readily generate confidence intervals about the following hypotheses, all of which satisfy the CE principle (Confidence Equivalence).

Setting
$b = 0$ yields hypotheses about μ_x (marginal)
$a = 0$ yields hypotheses about μ_y (marginal)
$0 \neq a = b$ yields hypotheses about $\mu_x + \mu_y$ (sums)
$0 < a = -b$ yields hypotheses about $\mu_x - \mu_y$ (differences)
$-1 = a, \quad b = \mu_x/\mu_y$ yields hypotheses about μ_x/μ_y (ratios)

It is interesting to note that for each of the first four problems, i.e. both marginals, sums and differences, the MST employs a calculable transformation, which also provides for a fiducial argument of the 'direct' type. Specifically, the 'natural' pivotals for these problems are:

$(\mu_x - x)/\sigma_x$ for the hypotheses about μ_x.

$(\mu_y - y)/\sigma_y$ for the hypotheses about μ_y.

$$\frac{[\mu_x + \mu_y] - [x + y]}{[\sigma_x^2 + \sigma_y^2 + 2\sigma_x\sigma_y\rho_{xy}]^{.5}} \text{ for the hypotheses about } (\mu_x + \mu_y)$$

$$\frac{(\mu_x - \mu_y) - (x - y)}{[\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\rho_{xy}]^{.5}} \text{ for the hypotheses about } (\mu_x - \mu_y).$$

However, the transformation for the fifth question, about ratios, is neither calculable nor does it induce a smoothly invertible pivotal. Therefore, there is no corresponding 'direct' fiducial method for hypotheses about μ_x/μ_y (see Appendix 9.8). The correct fiducial argument for hypotheses about the ratio is obtained through the step-by-step method, which yields a joint distribution for (μ_x, μ_y), and then integration over the appropriate cone-shaped area (centered at the origin) for probabilities about μ_x/μ_y.[21] That the confidence (Fieller's) solution was thought to be the correct fiducial answer is due to none other than Fisher himself. Before becoming aware of the conflict, he proposed the Fieller solution in both texts, *Statistical Methods for Research Workers* (§26.2) and *The Design of Experiments* (§62.1).[22]

A final instance of conflicting results between fiducial and confidence methods is type of the multiparameter problem where no Neyman-Pearson solution (nonmixed) exists; but, where a fiducial probability is available through the step-by-step method. An example of this, already discussed, is the Behrens-Fisher distribution. No additional analysis is needed for this case.

5.5. REVIEW OF PARADOXES: (2) CONFLICTS WITH BAYESIAN THEORY

Next, let us review the second category of difficulties, i.e. conflicts with Bayesian analyses of one sort or another. These conflicts divide naturally into two sub-headings: univariate and multivariate distributions.

The first of these has been rigorously examined in the preceding chapter. To summarize the findings there, Lindley's coherence results can be used since the restriction to 'natural' pivotals, i.e. smoothly invertible, canonical pivotals, guarantees the consistency of fiducial arguments for each of the first two univariate categories, i.e. for data of two kinds and for data of one-kind which can be combined into a single exhaustive summary. The third category, typified by Good's paradox (p. 130) has no universal solution; but, it was argued, the difficulty is general and is identifiable with the sensitivity of the Principle of Insufficient Reason to transformations of the unknown parameter. By insisting on a common sufficient statistic, the set of alternative formulations for the unknown parameter are limited to permit consistency.

The multivariate paradoxes are (with the exception of the univariate paradox, above), the most serious of all the challenges to the legitimacy of fiducial methods. Inconsistencies in fiducial inference are alleged, based on alternative fiducial techniques, akin to those discovered by Lindley and Good for one parameter distributions. Unfortunately, there is no multivariate generalization for Lindley's touchstone of coherence (as shown by Brillinger).[23] Therefore, the resolutions must all build on the intrinsic coherence of the step-by-step method, assuming it to introduce no inconsistency not already present in the single parameter case. The remedies fall under three titles, paralleling the three basic ways multivariate fiducial arguments can be abused. One – Failure to obey the step-by-step process through a 'direct' pivotal for several parameters simultaneously: examples being paradoxes of Mauldon, and Tukey.[24] Two – (the most common source) Failure to obey the 'naturalness' requirement, specifically canonicity, for the pivotals in both step_1 and step_3 of the step-by-step process; examples being paradoxes of Savage (in Tukey's article, above), and Brillinger (reference above). Three – (the most subtle error) Failure to obey the coherence check for the sufficiency claim, step_3; examples by Dempster, and Dawid, Stone and Zidek.[25] (This last article contains a very general criticism and will be briefly discussed at the conclusion of this section.)

Instead of dissecting each of the foregoing examples, I select representatives from each of the three classifications and offer resolutions to them. The others succumb to the same repartee. I am somewhat brief with the first two cases, since they shed little new insight into the fiducial argument. The third set of problems are philosophically relevant and deserve attention.

[1] The bivariate normal (p. 150) admits of a special factorization of Type O, when the correlation $\rho_{xy} = 0$. Assuming the variances equal to unity (for simplicity), we discover that:

$$\mathrm{p}(x, y/\mu_x, \mu_y) = \mathrm{p}(x/\mu_x) \cdot \mathrm{p}(y/\mu_y);$$

where each factor on the right hand side of the equality sign has a $\mathrm{N}(\mu, 1)$ distribution. Therefore, the smoothly invertible, canonical pivotals (w_1, w_2), defined as: $w_1 = (\mu_x - x)$ and $w_2 = (\mu_y - y)$; where each is $\mathrm{N}(0, 1)$, induce a joint fiducial posterior which is bivariate normal, with unit variance in all directions, centered about (x, y). (See the remarks on p. 150.)

Tukey uses the fact that this distribution is symmetric about the origin to generate a 'twisted' set of pivotals, defined as:

$$\begin{aligned} w_{1f} &= (\mu_x - x) \cdot \cos(f) + (\mu_y - y) \cdot \sin(f) \\ &= w_1 \cdot \cos(f) + w_2 \cdot \sin(f) \\ w_{2f} &= -(\mu_x - x) \cdot \sin(f) + (\mu_y - y) \cdot \cos(f) \\ &= -w_1 \cdot \sin(f) + w_2 \cdot \cos(f); \end{aligned}$$

where $f(\mu_x, \mu_y, r)$ is a sufficiently smooth function of three variables, and where $r = \sqrt{w_1^2 + w_2^2} = \sqrt{(\mu_x - x)^2 + (\mu_y - y)^2}$. That is, each circle of radius r is twisted through an angle $f(\mu_x, \mu_y, r)$. It follows that w_{1f} and w_{2f} are, for any pair of (μ_x, μ_y), distributed as (w_1, w_2), i.e. in a standardized bivariate normal distribution. Inverting on w_{1f} and w_{2f} does not induce the same posterior distribution for (μ_x, μ_y) as does the original pivotal set, even for the same datum point (x, y).[26]

The key to the response which removes this difficulty is the observation that f is not a calculable function unless both μ_x and μ_y are known. How does this affect the argument? The step-by-step method requires that, even for Type O distributions, each pivotal be 'natural' for a single parameter-variable pair. But the pivotals w_{1f} and w_{2f}, although smoothly invertible, fail to satisfy this requirement. The parameters for which these pivotals are defined are:

$$\mu_x \cdot \cos(f) + \mu_y \cdot \sin(f), \text{ for } w_{1f}, \text{ and}$$

$$-\mu_x \cdot \sin(f) + \mu_y \cdot \cos(f), \text{ for } w_{2f}.$$

The corresponding random variables for these pivotals are:

$$x \cdot \cos(f) + y \cdot \sin(f), \text{ for } w_{1f}, \text{ and}$$

$$-x \cdot \sin(f) + y \cdot \cos(f), \text{ for } w_{2f}.$$

However, since the parameters (μ_x, μ_y) are unknown, f is not a calculable function. Therefore, starting with data (x, y) the resulting random variables (above) are illegitimate replacements, being contaminated by the transformation f. Hence, the transformation for the original pivotals (w_1, w_2), although leading 1-1 to (w_{1f}, w_{2f}), is seen to be unsatisfactory when analyzed through the step-by-step method for each of the two transformed parameters. Mauldon's paradox makes use of the same sort of transformation to induce an invalid fiducial density, similar to that obtained from the pivotals (w_{1f}, w_{2f}).

[2] Brillinger's counterexample makes use of a bivariate exponential

distribution, defined as:

$$F(x, y/\alpha, \beta) = (1 - \exp[-\alpha x]) \cdot (1 - \exp[-\beta y]) \cdot (1 - \gamma \exp[-\phi x - \mu y]);$$

where x, y, α, β are all greater than 0; ϕ is a calculable function of α, μ is a calculable function of β, and γ is a known constant. First, note that the marginal distributions for (x, y) are exponential, given as:

$$F(x) = 1 - \exp[-\alpha x] \text{ and } F(y) = 1 - \exp[-\beta y].$$

Therefore, since a smoothly invertible, canonical pivotal exists for each of these distributions (see Example$_1$, p. 125), there is no problem with step$_3$ of the step-by-step method (assuming that the sufficiency claim can be verified by the coherence check).

The conditional distribution of, say y given x (needed for step$_1$), is:

$$F(y/x) = (1 - \exp[-\beta y]) \cdot (1 - \exp[-\phi x - \mu y] \cdot (1 - [\phi/\alpha] \cdot (\exp[\alpha x] - 1))).$$

Therefore, $F(x, y/\alpha, \beta)$ admits a factorization of Type *B* (see p. 149), with x and y reversed. In this respect, F is symmetric in x and y. That is, x and y can be reversed while maintaining a factorization of Type *B*. Using the step-by-step method with the different (but symmetric) factorizations leads to two posterior fiducial distributions which are mutually incoherent (assuming ϕ, μ and γ are determined as in Brillinger's paper).[27]

The resolution to Brillinger's paradox is to see that the fiducial argument is blocked at step$_1$. That is, using the conditional distribution $F(y/x)$ (for step$_1$), given x and α, there is no canonical pivotal for y and β. The conditional one-parameter distribution from step$_1$ fails to admit a natural pivotal for a fiducial argument from y to β. Brillinger's paradox reduces to a one-parameter problem, analogous to one of Lindley's distributions for which a coherent fiducial argument is unavailable. However, as has already been shown (p. 126), the reconstruction by smoothly invertible, canonical pivotals eliminates this difficulty.

Savage's example, although using a factorization of Type *C* (see p. 149) with similar symmetry for the two random variables, fails to adopt canonical pivotals in each of the two fiducial steps (1 and 3) and none are forthcoming.[28]

[3] The final violation of the step-by-step method to be examined consists

of the transgression of the check of coherence for the sufficiency claim in step_3 (see p. 139). This case has added significance as it points out the difference between the Objective-Bayesian approach and fiducial solution to inverse inference. The analysis given here suggests that the frequentist basis of the fiducial argument provides a satisfactory way out from some of the paradoxes of Dawid, Stone and Zidek.

A widely popularized paradox for the step-by-step method is due to Dempster. He reworks the step-by-step argument for the simple bivariate normal by switching from parameterization (μ, σ) to $(\sigma, \mu/\sigma)$. Under this formulation, the jointly sufficient statistics are $(s, \bar{x}/s)$, in place of the usual $(\bar{x}, s)$. The transformed distribution satisfies a factorization of Type *B*.

$$\mathrm{p}(s, \bar{x}/s \mid \sigma, \mu/\sigma) = \mathrm{p}(s \mid \sigma, \mu/\sigma, \bar{x}/s) \cdot \mathrm{p}(\bar{x}/s \mid \mu/\sigma).$$

Following Fisher's inverted step-by-step method, Dempster generates a step_3 fiducial distribution for (μ/σ) from $(\bar{x}/s)$. He also shows this fiducial probability to be identical to the one obtained by integration over the Fisherian joint distribution for (μ, σ). However, using step_1 of a fiducial argument for σ, given s, μ/σ and $\bar{x}/s$, he induces a marginal distribution for σ which differs from the one obtained with the usual pivotal from s to σ (see step_3, p. 143).[29]

There are several points of controversy in Dempster's alternative fiducial argument. First, he has not shown that the conditional pivotal of step_1 (from s to σ, given μ/σ and $\bar{x}/s$) is smoothly invertible. This is questionable since the conditional pivotal (of step_1) partitions the (μ, σ) plane in a fashion similar to that in which Fieller's faulty pivotal partitions the (μ_x, μ_y) plane, i.e. by lines radiating from a point.[30]

Second, and more important, the claim in step_3 that $\bar{x}/s$ is sufficient for μ/σ is valid when σ has the conjugate prior from the usual or Fisherian joint fiducial probability. This is obvious since, as Dempster shows, his fiducial posterior density for μ/σ given $\bar{x}/s$, agrees with Fisher's. But, the sufficiency claim in step_3 is accurate only in the absence of knowledge of σ (see p. 139). That ignorance is coherently represented by the usual conjugate prior for σ. Dempster's conjugate prior (from his step_1) is different and incompatible with the sufficiency claim. Just why this is so can be explained by considering the very important work of Dawid, Stone and Zidek.

Suppose that a bivariate distribution $\mathrm{p}(x, y/\theta_1, \theta_2)$ satisfies a factorization of Type *A*, with joint density:

$$\mathrm{p}(x/\theta_1, \theta_2) \cdot \mathrm{p}(y/\theta_2).$$

The authors have shown how it is possible to define an improper prior

probability $\pi(d\theta_1, d\theta_2)$ so that the posterior distribution for θ_2, i.e. $\mathrm{p}(\theta_2/x, y)$, is a function of y alone. But with the data y and corresponding density factor $\mathrm{p}(y/\theta_2)$, there is no conjugate prior for θ_2 (improper or not) which, through conditionalization, duplicates the original (marginal) probability for θ_2.

That this situation is paradoxical is made clear by the authors.[31] Using Bayes' theorem and CC, we have

$$\mathrm{p}(\theta_2/x, y) = \frac{\int_{\theta_1} \mathrm{p}(x, y/\theta_1, \theta_2) \cdot \pi(d\theta_1, d\theta_2)}{\iint \mathrm{p}(x, y/\theta_1, \theta_2) \cdot \pi(d\theta_1, d\theta_2)}$$
$$= \mathrm{g}(y, \theta_2, d\theta_2);$$

where, by assumption, g is some function of the three variables noted. Multiplying both sides of the right hand equations by the denominator yields:

$$\mathrm{p}(y/\theta_2) \cdot \int_{\theta_1} \mathrm{p}(x/y, \theta_1, \theta_2) \cdot \pi(d\theta_1, d\theta_2) =$$

$$\mathrm{g}(y, \theta_2, d\theta_2) \cdot \int\int \mathrm{p}(x, y/\theta_1, \theta_2) \cdot \pi(d\theta_1, d\theta_2).$$

If π is proper, i.e. if $\iint \pi(d\theta_1, d\theta_2) = 1$, then the last result may be integrated with respect to x, yielding:

$$\mathrm{p}(y/\theta_2) \cdot \pi(d\theta_2) = \mathrm{g}(y, \theta_2, d\theta_2) \cdot \int \mathrm{p}(y/\theta_2) \cdot \pi(d\theta_2);$$

where $\pi(d\theta_2) = \int_{\theta_1} \pi(d\theta_1, d\theta_2)$. But this last conclusion means that:

$$\mathrm{p}(\theta_2/x, y) = \mathrm{g}(y, \theta_2, d\theta_2) \propto \mathrm{p}(y/\theta_2) \cdot \pi(d\theta_2),$$

which is just the denial of the counterexample devised by the authors.[32]

The only promise for averting the Bayesian argument (above), hence for eliminating the paradox, is to avoid the replacement of $\int_{\theta_1} \pi(d\theta_1, d\theta_2)$ by $\pi(d\theta_2)$, in the last step (which uses the integration with respect to x). That is, since the prior probability for (θ_1, θ_2) is improper, the use of the integration technique for generating a 'marginal' prior must be blocked.

This is the suggestion offered by Dempster in his discussion of the paradoxes.[33] He argues that if the improper prior for (θ_1, θ_2) is given as the joint uniform improper one, i.e. $\pi(d\theta_1, d\theta_2) = d\theta_1 \cdot d\theta_2$, then the transformation to the equivalent parameter pair (η, ζ); where $\zeta = \theta_1$ and $\eta = \theta_1 \cdot \theta_2$, shows

up the inadequacy of the 'marginalization' technique. That is, the marginal for the first parameter, using the initial formulation (θ_1, θ_2), is just $d\theta_1$. However, under transformation, the joint prior density for (η, ζ) is just $d\eta \cdot d\zeta/|\zeta|$; from which the marginalization for ζ yields $d\zeta/|\zeta|$. However, $\zeta = \theta_1$. So by mere transformation, the prior 'marginal' for θ_1 has been changed from $d\theta_1$ to $d\theta_1/|\theta_1|$. Note the similarity to the transformation in Dempster's paradox for the fiducial step-by-step argument in the simple bivariate normal distribution. Parameters (μ, σ) were transformed to $(\sigma, \mu/\sigma)$ resulting in an abnormal density for σ. Perhaps, since there is a response to Dempster's paradox, there may be a retort to the Dawid, Stone and Zidek argument.

A resolution to some of these problems rests with the claim that y is sufficient for θ_2 in the absence of knowledge of θ_1. The coherence check in step_3 requires that, using the conjugate prior for θ_1 obtained from step_1, the fiducial argument from y to θ_2 via $\mathrm{p}(y/\theta_2)$ is consistent. Thus, if the distribution is one of the Dawid, Stone and Zidek sort where *no* conjugate prior for θ_2, conjoined with $\mathrm{p}(y/\theta_2)$, duplicates the original 'marginal' posterior distribution for θ_2, then the fiducial argument for step_3 is blocked and the paradox avoided. That is, no posterior joint fiducial density is defined. If the marginalization paradox for θ_2 can be avoided by using a special prior for θ_1, the fiducial step-by-step argument reflects this by permitting step_3 just in case the conjugate prior for θ_1 (from step_1) is the uniquely privileged one.

Dawid, Stone and Zidek, along with the discussants, interpret the paradoxes as a strong challenge to any program hoping to represent ignorance through improper prior probabilities. Thus, Objective Bayesians and fiducialists are brought under the gun. The first group adopt the improper priors outright. Once ignorance is appropriately formulated, the prior probability is specified. This results in improper 'ignorance' priors when the parameter has unbounded range. As for the second group, the fiducialists, they are committed (through appeal to Lindley's proof) to posterior probabilities which use exclusively conjugate priors of improper status. Therefore, argue the critics, the two programs fall to the Bayesian sword of the 'marginalization' paradoxes.

It has been shown here that by careful attention to the coherence check of step_3 some of the 'marginalization' paradoxes, as presented above, can be avoided. To repeat the results of our analysis, the coherence check is necessary as there is no multivariate counterpart to the one-dimensional sufficiency condition, expressed in terms of the stochastic properties of the

random variables used. Unlike the Objective Bayesian program, the commitment to conjugate 'ignorance' priors arises through the fiducial argument and not by any independent principle. Where the fiducial argument is blocked, e.g. by failure of the coherence check in $step_3$, there is no commitment to a prior probability which represents ignorance. Unlike the Objective Bayesian approach, which uses a modified Laplacean principle (with appeal to symmetries or subjective indifference to specify the ignorance prior), the fiducial argument relies on frequency based conditions to identify conjugate priors. This difference is reflected in the different responses to the marginalization paradoxes, with the advantage going to the Fiducialists.

Unfortunately, there remain serious deficiencies in the reconstruction offered here. With the exception of the last two problems (Dempster's and the marginalization paradoxes), there has been little evidence presented to cast doubt on the plausibility of using an *improper* prior distribution to represent ignorance. A corollary of Lindley's results on the necessary and sufficient conditions for coherence of fiducial inference with one parameter (and a sufficient statistic) is that, in location form, the conjugate prior is uniform.[34] Therefore, if the location parameter is unbounded the conjugate prior distribution is *improper*. In the two problems that follow, we shall discover paradoxical features of multivariate fiducial inference which are consequences of the impropriety of the conjugate priors.

Review, once again, the step-by-step fiducial argument for the normal distribution (pp. 141–143). Look, also, at Dempster's paradox (p. 159), where the parameters (μ, σ) are switched to the equivalent pair $(\sigma, \mu/\sigma)$. Dawid and Stone have investigated the posterior probability for these transformed parameters, given data $(\bar{x}, s)$, and based upon the conjugate, improper density $d\mu \cdot d(\ln \sigma)$.[35] This is also the conjugate prior for the step-by-step fiducial argument (but not the conjugate prior for Dempster's alternative solution).

Let $\theta = \mu/\sigma$ and $r = n\bar{x}/\sqrt{(n-1)s^2 + n\bar{x}^2}$, for n independent observations from $N(\mu, \sigma^2)$. Dawid and Stone have shown that the posterior probability density for θ, $p(\theta/\bar{x}, s)$, is proportional to

$$\exp\{-.5n\theta^2\} \cdot \int_0^\infty \omega^{n-2} \cdot \exp\{-.5\omega^2 + r\theta\omega\} d\omega. \qquad [*]$$

However, r has a distribution which depends upon θ only.

$$p(r/\mu, \sigma) = p(r/\theta) \propto \exp\{-.5n\theta^2\} \cdot (1 - r^2/n)^{\cdot 5(n-3)}$$

$$\cdot \int_0^\infty \omega^{n-1} \cdot \exp\{-.5\omega^2 + r\theta\omega\} d\omega. \qquad [**]$$

But, since [**] is *not* a factor of [*] as a function of θ, no prior distribution (improper or not) for θ will duplicate the posterior distribution [*] when it is conjoined with [**] through CC. Thus, using the conjugate prior sanctioned by the step-by-step method, we obtain a version of the Dawid, Stone and Zidek marginalization paradox for the transformed parameter θ. Specifically, the posterior distribution for θ depends upon the random variable r only. Also, r has a distribution dependent upon θ only. Nonetheless, one cannot rely on the statistic r for reaching the posterior distribution [*] by Bayes' theorem. If we call [**] the 'likelihood' function for θ given the variable r, then the posterior distribution for θ, given $(\bar{x}, s)$, is a function of r ([*]) but it cannot be duplicated by conjoining the 'likelihood' for θ, given r, with any prior probability for θ. The result is surprising, not contradictory.

A second and final illustration of the unexpected difficulties resulting from improper densities involves de Finetti's *Conglomerability* principle (see Appendix 9.1 for discussion of its connection to fiducial inference). Conglomerability is defined in the following way. Suppose that an event H can be partitioned into pairwise exclusive and mutually exhaustive H_α, with α ranging over all cardinals less than some cardinal β. Let the conditional probability (density) $\mathrm{p}(E/H_\alpha)$ satisfy the inequality

$$k_1 \leqslant \mathrm{p}(E/H_\alpha) \leqslant k_2, \text{ for all } \alpha.$$

Then conglomerability asserts that

$$k_1 \leqslant \mathrm{p}(E/H) \leqslant k_2.$$

With a *proper* probability for the elements H_α conglomerability follows (see Appendix 9.1). Such is not the case with *improper* priors and the results prove embarassing for fiducial inference.[36]

A simple example of this phenomenon is based on the work of R.J. Buehler and A.P. Feddersen.[37] They have shown that with a sample of two independent trials (x_1, x_2) from the normal distribution $\mathrm{N}(\mu, \sigma^2)$, the following inequality holds for all values of μ and σ^2.

$$\mathrm{p}(x_{\min} \leqslant \mu \leqslant x_{\max} / [|t| \leqslant 1.5], \mu, \sigma^2) > .518, \qquad [\dagger]$$

$t = |x_1 + x_2|/|x_1 - x_2|$ and $x_{\min}$ ($x_{\max}$) is the minimum (maximum) of (x_1, x_2).

Now, we have already seen (p. 143) that the marginal posterior distribution

for μ, given (x_1, x_2), obtained with the step-by-step method, is the familiar t-distribution centered on $\bar{x}$, with 1 degree of freedom. This is a Cauchy distribution with quartiles at $x_{\min}$ and $x_{\max}$. Hence,

$$p(x_{\min} \leqslant \mu \leqslant x_{\max}/(x_1, x_2)) = .5 \qquad [\dagger\dagger]$$

given any data (x_1, x_2). Since the event $|t| \leqslant 1.5$ can be partitioned by the set of (x_1, x_2) comprising it, by conglomerability

$$p(x_{\min} \leqslant \mu \leqslant x_{\max}/[|t| \leqslant 1.5]) = .5. \qquad [\dagger\dagger\dagger]$$

Since [†] holds for any values of μ and σ^2, by conglomerability

$$p(x_{\min} \leqslant \mu \leqslant x_{\max}/[|t| \leqslant 1.5]) > .518. \qquad [\dagger\dagger\dagger\dagger]$$

The formulas [†††] and [††††] conflict. In fact there is a way out of the problem without admitting a contradiction. Conglomerability is violated since the probability

$$p(x_{\min} \leqslant \mu \leqslant x_{\max}/[|t| \leqslant 1.5])$$

is *not defined* if the improper conjugate prior $d\mu \cdot d(\ln \sigma)$ is used. Thus the paradox can be avoided but only at the expense of conglomerability.[38]

The last two problems point to the need for additional examination of standard inductive principles when improper probability distributions are allowed. One can note, with a sigh, that the burden for this examination is carried jointly by Fiducialists and Objective Bayesians.

NOTES

[1] Fisher, *Statistical Methods and Scientific Inference*, p. 179.

[2] As demonstrated in the discrete example that follows, the 'marginal' fiducial probability for θ_1, generated by CF, may conflict with Neyman-Pearson confidence intervals for θ_1. As could be expected, the discord stems from the attempt to use a 'repeated sampling' interpretation of the marginal fiducial probability.

[3] M.S. Bartlett, 'Statistical Information and Properties of Sufficiency', *Proc. Royal of London, Ser. A.* 154 (1936), 124–137.

[4] G. Barnard, 'Some Logical Aspects of the Fiducial Argument', *J. Royal Stat. Soc., Ser. B* 25 (1963), 111–4; J.D. Kalbfleisch and D.A. Sprott, 'Application of Likelihood Methods to Models Involving Large Numbers of Parameters', *J. Royal Stat. Soc., Ser. B* 32 (1970), 175–208; A.P. Dawid, M. Stone and J.V. Zidek, 'Marginalization Paradoxes in Bayesian and Structural Inference', *J. Royal Stat. Soc., Ser. B* 35, #2 (1973), 189–233.

[5] See the discrete example that follows on page 143 for an illustration of this phenomenon.

[6] Jeffreys warns readers of the problem and even shows its connection to group

theoretic invariance procedures (Fraser's theory), but he offers no solution. Jeffreys, *Theory of Probability*, pp. 191–192.

[7] Fisher, *Statistical Methods and Scientific Inference*, pp. 123–124.

[8] *Ibid.*, pp. 82–84.

[9] The 'direct' fiducial argument is not to be confused with the simultaneous fiducial method (page 120) using a sufficient (or exhaustive) statistic for combining data in one parameter problems.

[10] *Ibid.*, pp. 123–4. In the above remarks, S, the 'sum of squares' equals $s^2 \cdot (n-1)$; where s^2 is the 'estimated sample variance.'

[11] *Ibid.* In the above, read 'simultaneous distribution' as synonymous for 'joint distribution.'

[12] D.V. Lindley, 'Review of Statistical Methods and Scientific Inference', *Heredity* 11 (1957), 280–3; see p. 281.

[13] Fisher, 'On Some Extensions of Bayesian Inference Proposed by Mr. Lindley', *J. Royal Stat. Soc., Ser. B* 22 (1960), 299–301; F. Yates, 'Comments on D.V. Lindley's Review of Sir Ronald A. Fisher's *Statistical Methods and Scientific Inference*', *Heredity* 12 (1958), 133–5.

[14] Fisher, 'The Comparison of Samples with Possibly Unequal Variances', *Annals of Eugenics* ix, 174–80. Rep. in *CMS*, paper #35.

[15] Fisher, *Statistical Methods and Scientific Inference*, pp. 175–80. Also, for some interesting commentry see: G.W. Bennett and E.A. Cornish, 'A Comparison of the Simultaneous Fiducial Distributions Derived from the Multivariate Normal Distribution', Conference on Fiducial Probability, *Bull. Int. Stat. Inst.*, 40^2 (1963), 902–18.

[16] See Appendix 9.7 for details.

[17] See Appendix 9.7 for details.

[18] Fisher, *op. cit.* Fisher skips the next to last step, going directly to a joint distribution for the variances, given the correlation. Details of the rigorous step-by-step inference for these stages are found in 'Sampling the Reference Set', *Sankhya* 23 (1961), 3–8. Fisher uses $\rho_{xy} = 0$. The marginal distributions for the population means and variances are *not* the familiar 'Student' and χ^2 distributions. See, S. Geisser, 'Book Review of *Introduction to Probability and Statistics from a Bayesian Viewpoint*', *Technometrics* 8, #2 (May, 1966).

[19] B.L. Welch, 'On Confidence Regions and Sufficiency, with Particular Reference to Parameters of Location', *Annals Math. Stat.* 10 (1939), 58–69. Discussed in D.A.S. Fraser, 'On Fiducial Inference', *Annals Math. Stat.* 32, #3 (1961), 661–76; see p. 671.

[20] Fisher, *op. cit.*, pp. 101–2.

[21] This last procedure corresponds to the method suggested by M. Creasy in her article, M. Creasy, 'Limits for the Ratio of Means', Symposium on Interval Estimation, *J. Royal Stat. Soc., Ser. B* 16 (1954), 186–94.

[22] There is a minor difference in Fisher's formulation since he considers the problem with the variance ratio known to be 1, but the variances unknown. His actual solution, therefore, is the Student-ized version of Fieller's test.

[23] D.R. Brillinger, 'Examples Bearing on the Definition of Fiducial Probability with a Bibliography', *Annals Math. Stat.* 33 (1962), 1349–55.

[24] J.G. Mauldon, 'Pivotal Quantities for Wishart's and Related Distributions, and a Paradox in Fiducial Theory', *J. Royal Stat. Soc., Ser. B* 17 (1955), 79–85; J.W. Tukey, 'Some Examples with Fiducial Relevance', *Annals Math. Stat.* 28 (1957), 687–95.

[25] A.P. Dempster, 'Further Examples of Inconsistenices in the Fiducial Argument', *Annals of Math. Stat.* 34 (1963), 884–91; A.P. Dawid, M. Stone and J.V. Zidek, 'Marginalization Paradoxes in Bayesian and Structural Inference', *J. Royal Stat. Soc., Ser. B* 35 #2 (1973), 189–233.

[26] Tukey, *op. cit.*, p. 690.

[27] Brillinger, *op. cit.*, p. 1352.

[28] Brillinger proposes an extension of Lindley's coherence results for the one-parameter case, and then constructs a counterexample to demonstrate that the proposed conditions are not necessary (although they may be sufficient) for a two-parameter, coherent fiducial argument, *Ibid.*, pp. 1352–3.

But the step-by-step argument is not restricted by Brillinger's suggested modification of Lindley's conditions.

Brillinger's proposed extension of Lindley's results is that: (for a two-parameter case) There exists transformations of:

$$x \text{ to } a(x, y, \mu, \nu) \qquad y \text{ to } b(x, y, \mu, \nu)$$

and

$$\alpha \text{ to } c(\alpha, \beta, \mu, \nu) \qquad \beta \text{ to } d(\alpha, \beta, \mu, \nu)$$

forming a group for arbitrary parameters μ and ν, and such that the c.d.f. satisfies: $F(x, y, \alpha, \beta) = F(a, b, c, d)$. These are, of course, alleged to be both necessary and sufficient conditions for a coherent, two-parameter fiducial posterior distribution.

Brillinger's counterexample for the necessity claim of his own conjecture is found in the bivariate distribution, given as:

$$F(x, y, \alpha, \beta) = 1/((1 + \exp[-x - \alpha]) \cdot (1 + \exp[-\phi(\alpha) \cdot (y + \beta)]));$$

where ϕ is a calculable function of α. As he shows, this distribution admits of a joint fiducial posterior density with compatible Bayesian conjugate priors (which are the uniform 'ignorance' ones). But, F is *not* invariant under any two parameter transformation.

The natural pivotals for this distribution are:

$$v_\alpha = (\alpha + x) \text{ and } v_\beta = (\beta + y);$$

which are canonical with $\alpha^* = 0$ and $\beta^* = 0$. This distribution satisfies a factorization of Type *A*. Hence, the step-by-step multivariate fiducial argument, using canonical pivotals, is not restricted to Brillinger's proposed extension of Lindley's results.

[29] Dempster, *op. cit.*, pp. 884–7. Also, his step_3 pivotal is not canonical.

[30] *Ibid.*, p. 887.

[31] A.P. Dawid, M. Stone and J.V. Zidek, *op. cit.*, p. 194.

[32] *Ibid.*, p. 194.

[33] Dempster, Discussion of Dawid, Stone and Zidek paper, *op. cit.*, pp. 222–3.

[34] D.V. Lindley, 'Fiducial Distributions and Bayes' Theorem', *J. Royal Stat. Soc., Ser. B* 20–21 (1958), 102–7.

[35] A.P. Dawid and M. Stone, 'Un-Bayesian Implications of Improper Bayes Inference in Rountine Statistical Problems', *Biometrika*, 59 #2 (1972), 369–75.

[36] The idea of showing a tacit assumption of conglomerability underlying some of the 'paradoxes' is a consequence of I. Levi's illuminating view that, since the use of *improper* distributions entials a failure of additivity, conglomerability is not a valid inference form.

[37] R.J. Buehler and A.P. Feddersen, 'Note on a Conditional Property of Student's *t*', *Annals Math. Stat.* 34 (1963), 1098–100.

[38] C. Stein has constructed a problem in which fiducial and confidence solutions differ widely, 'An Example of Wide Discrepancy Between Fiducial and Confidence Intervals', *Annals Math. Stat.* 30 (1959), 877–80. What is most striking about Stein's example is that the two solutions are mutually incompatible of conglomerability is assumed. That is, like the Buehler-Feddersen 'paradox' discussed in the text, not only does the fiducial solution fail to support a repeated sampling interpretation, but with a suitable partition of the unknown states, certain sets of inequalities are inconsistent if conglomerability is allowed.

Specifically, Stein's ingenious construction involves a sequence of independent random variables $(x_1, \ldots, x_n)$ from the normal distributions $N(\mu_i, 1), i = 1, \ldots, n$. Inference is with respect to the sum of the squares of the unknown means, $\Sigma_i \mu_i^2$. The fiducial solution is based on a multivariate normal posterior distribution, with means $(x_1, \ldots, x_n)$. The confidence solution is derived from $\Sigma_i x_i^2$, which has a non-central χ^2 distribution, with n degrees of freedom and non-centrality parameter $\Sigma_i \mu_i^2$. Stein shows that if $lim_{n\to\infty}(1/n^2)\Sigma_i \mu_i^2 = 0$, then, for any value of $\Sigma_i \mu_i^2$ and n sufficiently large, the upper α fiducial probability interval for $\Sigma_i \mu_i^2$ covers the true value with probability approaching 0. The appeal to conglomerability is made in shifting from the inequality for coverage probability (conditional on $\Sigma_i \mu_i^2$) to the fiducial probability (conditional on $(x_1, \ldots, x_n)$. As in the Buehler-Feddersen example, an analogue to [†††] is not defined. That is, I believe, conglomerability fails because probability for the 'coverage,' conditional upon $\Sigma_i \mu_i^2/n^2 < \epsilon$, is not defined from the *improper* prior distribution $\Pi_i d\mu_i$.

Finally, note that the step-by-step method requires a logically weaker principle than conglomerability. In terms of L. Dubins' analysis (see his 'Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations', *Annals of Probability* 3, #1 (1975), 89–99), all we require is π-disintegrability (or equivalently, π-conglomerability) for *some* partition π.

CHAPTER 6

IAN HACKING'S THEORY

In recent years two philosophers of distinguished merit, I. Hacking and H. Kyburg, have offered thorough criticism of contemporary statistical practice. In particular, each has presented a novel reconstruction of Fisherian inference, with special attention to the fiducial argument. In this book, I am concerned with inverse inference generally, and fiducial probability specifically. The preceding chapters have shown where the outstanding difficulties lie for Fisherian inference. In the upcoming chapters, I hope to connect the analysis of fiducial inference with a discussion of more general inverse inference by considering relevance principles. The theories advocated by Hacking and Kyburg differ on rules governing relevant evidence in ways that point out the central problems for integrating fiducial theory with inference of a general sort. More to the point, the text investigates the compatibility of fiducial inference with such common practices as randomization in experimental design, (6.2, 7.2 and 8.2).

Hacking's theory is examined in Chapter 6. In 6.1 his views on fiducial inference are presented. In 6.2, I compare these with his recommendations for inverse inference based on overt randomization.

Hacking's reconstruction of the fiducial argument rests on his novel principle of irrelevance, which is crucial for justifying the claim that the data are irrelevant to probability statements about the pivotal variable. (See 4.2.) Cast as a *likelihood principle of irrelevance*, Hacking's rule serves to link fiducial and likelihood concepts. However, because of a special technical device, *experimental density*, the tie to Fisherian likelihood is strained. Hacking's version of fiducial inference satisfies coherence conditions (in one parameter problems) up to the level of Lindley's analysis. As with the reconstruction offered in Chapter 4, the problem discussed in 4.8 is not avoided by Hacking's principle.

In a fiducial argument, one tries to show how the sample data are typical of the population from which they come. The pivotal variable is the formal mode for characterizing the 'representativeness' of the sample. In inverse inference one must evaluate the representative features of the data quite generally. A standard methodological tool for enhancing the representative status of a sample (for inverse inference) is the device of randomization: picking items to be sampled with the aid of a randomizer. Hacking's version of fiducial inference (as well as that of Chapters 4 and 5) requires a *demonstration of irrelevance* before the crucial, 'after trial' step can be made. But with randomization, it appears that *relevance must be demonstrated*, if the benefits of randomization are voided This shift in the attitude taken towards relevance conditions is explored in Section 6.2.

6.1. IAN HACKING'S LIKELIHOOD PRINCIPLE OF IRRELEVANCE AND THE FIDUCIAL ARGUMENT

It is with respect to both the understanding of the fiducial argument and the general concern for a clarification of inductive relevance that Ian Hacking's

work, *Logic of Statistical Inference*[1] becomes centrally important. His claim is no less than to have discovered a novel principle of relevance which can be used as the lynch pin in a reconstruction of the fiducial argument: a reconstruction in accord with frequentist doctrines. What makes Hacking's approach alluring is that when frequency based probabilities are not available, the inductive support is defined by likelihood. Thus, Hacking uses only resources recommended by Fisher.

The modest goal of this section is to examine Hacking's reconstruction and show that, although compatible with the arguments of the previous sections, it fails to answer any questions not already responded to. On the other hand, Hacking's work is most elegant and completes the inductive hierarchy of Fisherian methods. Fisher used three forms of quantitative inductive inference: significance, likelihood and fiducial probability. They represent, in increasing order, more sophisticated concepts of inductive support. The previous chapters have attempted to link the first and third of these methods. Hacking, through use of his novel principle of irrelevance, links the second and third. In addition, Hacking's reconstruction can claim to be the first (historically) which supplies precise conditions for the fiducial argument, as well as a clear picture of how the argument proceeds.

A good starting place for a review of Hacking's position is with the observation that he implicitly accepts a dual probability semantics. He limits the domain of his logic of support to cases where the background knowledge contains frequency or 'chance' information. For Hacking, a statement can serve as evidence, or for that matter as a hypothesis, only if it is statistical.

The data must be included, in Hacking's jargon, in a 'joint proposition' if it is to be counted as evidence. Joint propositions are constructed from six elements $\langle X, K, D; T, K', E\rangle$. First, there must be a 'chance set-up' (X). Examples of chance set-ups include coins and coin flipping devices, or urns containing counters. It is clear that the chance set-up is a part of the real world, but precisely which part is not obvious. Does it include the investigator running the chance device, the table supporting the device, the room containing the urn? No rule is given, and the question remains open.

Second, the data must refer to a 'kind of trial' (K) conducted on the chance set-up. Hacking is not explicit in his definition of a kind of trial, but from his examples one can extract two ingredients. The kind of trial must designate how the chance set-up is to be used, e.g. flip the coin twice, draw from the urn until a green counter shows. Also, it must dictate how the outcome of the trial on the chance set-up is to be described (at least initially),

e.g. note the order of the outcomes of the two flips; record how many counters were drawn. The first ingredient in the kind of trial is called the *sampling rule*, that is, the rule which states how the outcome of the trial is reached. The second ingredient, the description of the outcome, is called the *statistic*. The combination of the sampling rule and statistic defines the set of possible outcomes of trials on the given chance set-up. This set is called the *sample space* (Ω in Hacking's jargon).

The third element in the joint proposition is the frequency distribution (D). It stipulates what the frequency, 'chance' is for each outcome in the sample space, given the set-up and kind of trial. To use our familiar binomial example: if the chance set-up is a coin and flipping device, and the trials are defined by the sampling rule and statistic which state: flip twice and note the number of heads showing; then, since the sample space consists of the three outcomes $\{0, 1, 2\}$, a binomial distribution D corresponding to the hypothesis that the coin is fair assigns the frequencies $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$ (respectively). Alternative hypotheses about the bias of the coin, using the same set-up and sample space, are represented by the set of distributions $\Delta = \{(1-\theta)^2, 2 \cdot \theta \cdot (1-\theta), \theta^2\}$. Here, θ is a binomial parameter interpreted as a measure of the bias of the coin.

The fourth, fifth and sixth elements in the joint proposition combine to assert that a particular instance T of trial K' (not necessarily the same as K – the statistic for K' might be a reduction of the one for K) yields the specified outcome E. To continue with the coin example, the joint proposition might be completed by stating that the next two flips of the coin result in a pair of 'heads'.

If the background knowledge is strong enough to single out a unique distribution (as opposed to a family of distributions, e.g. Δ), then Hacking calls the resulting joint proposition *simple*. Two joint propositions are said to be *equivalent* if their component propositions are equivalent, relative to the background knowledge. Finally, a joint proposition e *includes* the joint proposition h: $\langle X, K, \Delta; T, K', \Omega\rangle$ iff e is equivalent to a joint proposition e': $\langle X, K, \Delta'; T, K', \Omega'\rangle$; where $\Delta \subseteq \Delta'$ and $\Omega \subseteq \Omega'$.

Hacking's support function is generated by the 'likelihood' for simple joint propositions. If j is a simple joint proposition, then the likelihood of j is the chance (frequency probability$_2$) of getting outcome E on trial T of kind K', given D. For instance, in the fair coin example, let j include the statement that a pair of 'heads' will occur on the next two flips of this *fair* coin. Then j has likelihood $\frac{1}{4}$. Likelihood is used as a measure of relative support for two simple joint propositions if there is a common evidential base for comparison.

The Law of Likelihood: If h and i are two simple joint propositions included in a joint proposition e, then given evidence e (alone), h is better supported than i if the likelihood of h exceeds the likelihood of i.[2]

The Law of Likelihood combines under one heading what has traditionally been considered two separate forms of inference. The hypothesis being supported can surpass the deductive content of its evidential basis in either of two ways. First, it can propose the occurrence of a particular outcome whereas the joint proposition used as evidence states only a disjunction of possible outcomes. A typical case arises when the entire sample space Ω is the evidential basis.

Secondly, the hypothesis can propose a simple distribution D in place of a set of possible distributions. An extreme, but common case occurs when the evidence specifies only a large family of distributions, called a statistical model. For example, to repeat the coin flipping problem, the joint proposition used as supporting evidence cites the whole set Δ of binomial distributions. That is, the background knowledge shows only that the chance set-up is binomial.

Instances of the first sort are labeled problems of *direct* inference; from a known distribution to a hypothesized outcome. The latter cases are referred to as problems of *inverse* inference; from known outcomes to postulated distributions. Hacking's use of likelihood to cover both problems marks a break with classical procedures. Fisher, who is of course responsible for the use of the likelihood function, considered it a one-place function of distributions for fixed outcomes. Thus, Fisher, and in this respect many contemporary statisticians, use likelihood for inverse inference only. Hacking however, uses it as a two-place function of both outcomes and distributions.

The broad outlines of Hacking's strategy for reconstructing the fiducial argument, and also the outlines of a connection between his reconstruction and that proposed here, are present in the foregoing analysis pointing out the double edged aspect of Hacking's likelihood function. Fisherian significance was presented as a problem of *direct inference,* where predictions used the D_p discrepancy measure for ranking the members of the sample space. Hacking's Law of Likelihood, applied to the same problem, leads to the very same ranking over Ω, given h_0. In Chapter 4, it was suggested that a reconstruction of fiducial inference could be based on *direct inference,* involving statements about the pivotal *both before and after* seeing the random variable.

The crucial amendment to Fisher's restrictions (1) and (2) was (3) canonicity (Section 4.6). It was justified by linking up statements of significance with those of fiducial probability (under location parameter form)

and showing that the two are consistent. Hacking, on the other hand, replaces (1)–(3) with his own likelihood rule; now interpreting likelihood as a function of the data (in the traditional way) *and* in his own special way (appropriate for direct inference). Thus, by comparing statements about the pivotal based on the second interpretation of likelihood (*without* knowing the experimental outcome) to those statements about the pivotal based on the first interpretation of likelihood (*after* observing the *r.v.*), Hacking is able to replace (1)–(3) with a novel Likelihood Principle of Irrelevance, and thereby achieve the desired results.

Hacking's proposal to use likelihood to obtain probabilistic support for direct inference is formalized in the frequency principle$_1$:

> *Frequency Principle*$_1$: If '$\mathbf{P}_D(E)$' is the (frequency) chance of outcome E according to the distribution D, then,
>
> $$\mathbf{p}(\langle X, K, D; T, K, E\rangle / \langle X, K, D; T, K, \Omega\rangle) = \mathbf{P}_D(E).^3$$

It says, simply, that if the chance (probability$_2$, $\mathbf{P}$) for outcome E is given by the evidence (a joint proposition which includes a distribution D for specified trials on a given set-up), then the inductive support (a probability$_1$, $\mathbf{p}$) for the hypothesis that E occurs as the outcome on a designated trial of the sort defined in the evidential base, is just the frequency value supplied by the distribution D. By restricting the evidence to a single joint proposition specifying not only the distribution, but also the set-up and kind of trial, Hacking circumvents what, for frequentists, is the primary problem of relevance for direct inference, i.e. the reference class problem.

An important development in Hacking's theory is the use of likelihood to obtain probability support for hypotheses, in cases of direct inference, even when there is no unique distribution posited in the evidential basis. Suppose the evidence e specifies only the chance set-up, kind of trial and set of possible distributions Δ. Suppose also, that for each distinct distribution $D \in \Delta$ the chance of a specific outcome $E^* \subseteq \Omega$ is constant, say k. Then, Hacking generalizes his frequency rule to include the result that the support for the hypothesis that E^* occurs, given only the evidence e, is k.

> *The Frequency Principle*$_2$: If '$\mathbf{P}_\Delta$' is the chance for E^* (a constant value k, by hypothesis), then,
>
> $$\mathbf{p}(\langle X, K, \Delta; T, K', E^*\rangle / \langle X, K, \Delta; T, K', \Omega\rangle) = \mathbf{P}_\Delta(E^*).^4$$

The novelty of the second frequency principle in a theory of support founded on likelihood is made evident when it is remembered that traditionally,

likelihood is a function of the distributional part of the hypothesis. Thus, the likelihood of a proposition with a disjunctive distributional part is not defined. It is the counterpart to a composite null hypothesis.

For example, if a coin has been flipped and the observation is: a 'head', then even though the likelihood for the hypothesis that the coin is fair ($\theta = .5$) is just .5, and the likelihood for the hypothesis that the coin is biased (say, $\theta = .8$) is just .8, the likelihood for the disjunctive hypothesis that the coin is fair or biased is *not* defined. It is very much like asking, what is the income of Peter or Paul? To answer, one must know which one. Hacking's second frequency rule permits an answer if both have the same income. Thus, it corresponds to the simple D_p ranking for exact significance tests (for composite h_0). Of greater importance is that the second frequency principle is justifiable by an appeal to CC and Bayes' theorem (if only proper distributions are admitted).[5]

Turning next to the problem of inverse inference and the fiducial argument, Hacking's reconstruction hinges on a transformation to direct inference. One pin in the argument is the elimination of the prior probabilities for the distributions D's $\in \Omega$ through the use of the second frequency principle. The other pin in the argument is the novel Likelihood Principle of Irrelevance. With the addition of two new terms, a *logical field* and a *partition*, the novel principle can be given.

DEFINITION: a *partition* of the joint proposition e is a set of mutually exclusive propositions such that e implies some member of the set is true.

DEFINITION: A *logical field* based on a partition A of e is the smallest set of propositions closed under the operations of (countable) alternation and negation with respect to e, and which includes every member of the partition A.

[LPI] *Likelihood Principle of Irrelevance:* If A is a partition of both joint propositions d and $d\&e$, while the likelihood of each member of A, given d, equals its likelihood given $d\&e$, then e is said to be *irrelevant* to A, given d.[6]

The Likelihood Principle of Irrelevance [LPI] (distinct from [LIP], p. 119), in conjunction with the frequency principles, leads to the probabilistic principle of irrelevance.

[PPI] *Probabilistic Principle of Irrelevance:* If e is irrelevant to A given d, and h is in the logical field based on A, then

$$p(h/d) = p(h/d \,\&\, e).^{7}$$

The focus of Hacking's reconstruction, using PPI (through LPI) and the Frequency Principle$_2$, is the pivotal variable; which is an (unobserved) outcome of a derived kind of trial. The procedure is best illustrated by Hacking's own, discrete example (p. 108). Again, given the restricted *BK* which includes only the binomial assumption, and the knowledge that $\theta = .4$ or $\theta = .6$ (without frequency 'priors' for the two values), the pivotal variable o is defined as:

$$o = \begin{array}{l} o_1 \text{ iff ('heads' and } \theta = .6) \text{ or ('tails' and } \theta = .4) \\ o_2 \text{ iff ('tails' and } \theta = .6) \text{ or ('heads' and } \theta = .4) \end{array}$$

Given just *BK* and using the Frequency Principle$_2$, the support for the hypothesis $h_{o_1}: o_1$ occurs, is .6; and the support for the hypothesis $h_{o_2}: o_2$ occurs, is .4. Note that these support statements are probabilities. Consider what happens with the addition of evidence e: The outcome of the flip is a 'head'. Hacking argues that *if* e is irrelevant to the hypotheses h_{o_1} and h_{o_2} (given *BK*), then since given e h_{o_1} is deductively equivalent to $h_1: \theta = .6$, and h_{o_2} is deductively equivalent to $h_2: \theta = .4$, probabalistic support for hypotheses about θ is available. That is, LPI and PPI can be used to detach the consequence, by granting the antecedent that e is irrelevant to h_{o_1} and h_{o_2}.

Looking back at the reconstruction offered in Chapter 4, we see that Hacking's argument follows the same path, but supplies different grounds for detaching the Fiducial Step. Instead of the pivotal restrictions: (1) existence, (2) smooth invertibility and (3) canonicity; Hacking uses his frequency principles and LPI. (See Appendix 9.6 for a discussion of Hacking's Frequency Principle$_2$.) How do Hacking's conditions compare to the three pivotal restrictions?

The answer is rather surprising. In order for Hacking to use the Frequency Principle$_2$ to get probabilistic support (not merely likelihood support) for outcomes of the pivotal variable, the pivotal must have a distribution independent of the unknown parameter. Thus, (1) existence of the pivotal is presupposed in Hacking's reconstruction. Second, the antecedent of LPI requires a common partition both before and after the data are observed. As demonstrated in Appendix 9.5, Tukey's smooth invertibility condition follows from LPI; notably, invariance of the partition to observed data follows from condition (*A*) of smooth invertibility (see p. 111).[8] Finally, the residual content

of LPI is traded off against condition (3) canonicity. Without attempting a direct comparison of canonicity to LPI we can assess the effects of Hacking's LPI by reviewing how his reconstruction fares against the three conditions of adequacy (proposed on p. 125).

With respect to the first one, coherence, Hacking argues (pp. 154–155 of his book) that any one parameter fiducial argument which satisfies his LPI condition also satisfies Lindley's conditions for coherent fiducial inference. The second condition of adequacy also is not violated since LPI does not rely on an acceptance of unwarranted prior probabilities to establish coherence. Of course, Lindley's arguments verify that a unique conjugate prior, namely the (improper) uniform 'ignorance' prior (for location parameter form), can be used to demonstrate the coherence.

There remains the final challenge, to show a tie between the new irrelevance principle and other Fisherian doctrines, to guard against any charge of *ad-hocery*. We have already discussed how LPI makes use of likelihood based support for appraising the inductive relevance of the observed data to hypotheses about the pivotal variable. In the example that follows this connection is investigated once again, this time by considering a fiducial argument with a continuous random variable. The purpose of using a continuous variable is to emphasize a different aspect of Hacking's reconstruction, an aspect which ensures the limited coherence offered by Lindley's results.

Reexamine the use of LPI in the discrete example (p. 174). The core of the principle is that the likelihood for pivotal outcomes must be the same *before* and *after* observing the flip of the coin. Ignoring the partition (A), we know that the pivotal variable o has an initial chance distribution $P(o_1) = .6$ and $P(o_2) = .4$. Therefore, using just the Law of Likelihood (p. 171), we can assign levels of inductive support to hypotheses about the pivotal variable. However, when LPI is used to defend the irrelevance claim (needed for the Fiducial Step), likelihood is determined with repsect to the elements of the partition. This is not the same procedure used in the Law of Likelihood. In discrete distributions this use of the partition leads to the same inductive support for pivotal outcomes as the support derived from the Law of Likelihood. The agreement in continuous distributions is acquired (quite cleverly) by restricting the random variables used to calculate likelihoods.

Hacking's solution bases likelihood on *experimental density*. Suppose that an experimental set-up yields observations which are continuously distributed according to the exponential law:

$$F(x, \theta) = 1 - \exp[-x\theta],$$

as, for example, in Fisher's emission count problem on p. 117. The cumulative distribution function [c.d.f.] $F(x, \theta)$ gives the probability of getting an experimental reading less than or equal to any prescribed value, say x_0, for a given value of θ. Probabilities for exact, point valued outcomes are literally zero; however the density fucntion $\partial F/\partial x$, defined as

$$\lim_{\epsilon \to 0} \frac{F(x+\epsilon, \theta) - F(x, \theta)}{\epsilon} = f(x, \theta)$$

permits the comparison of chances for different point values outcomes of x through the ratio of experimental densities, i.e.

$$f(x_0, \theta)/f(x_1, \theta) = \lim_{\epsilon \to 0} \frac{F(x_0+\epsilon, \theta) - F(x_0, \theta)}{F(x_1+\epsilon, \theta) - F(x_1, \theta)}$$

Experimental densities are defined as the density function with respect to a particular random variable. The reason for the designated form is simply that the density function is not invariant to transformations of the variable, even for calculable ones. For example, if $T: x \to x^2$ is used to switch variable then $f(x, \theta) \neq f(x^2, \theta)$, generally.[9]

Hacking gives the following rule for selecting the random variable used to calculate experimental densities.

> So we shall insist that the density function for defining likelihood ratios should be the *experimental density*. This is to be defined as the derivative of the cumulative [distribution] function with respect to the quantities actually being measured.[10]

Thus, experimental densities are defined in terms of the kind of trial to be run, which in turn, specifies the actual quantity to be measured.

For example, suppose that x_F (the value of cumulative distribution function) is to serve as the pivotal variable. Even though it is never 'observed' (since θ is not known), the Law of Likelihood can induce a support function for statements of direct inference about x_F. However, since the likelihood is calculated with repsect to the variable to be measured (x), the density function for x_F is *not* the appropriate one to use. That is, even though x_F is known to be uniformly distributed on the interval [0, 1], based on the *BK* and *before* seeing x, the Law of Likelihood may assign unequal support to hypotheses about the value of F.

For instance, in the exponential distribution:

$$F(x, \theta) = 1 - \exp[-x\theta] = x_F$$

the likelihood for the hypothesis that $x_F = .5$ is calculated with respect to x;

hence equals:

$$x \cdot \exp[-x\theta] = x \cdot (1 - x_F) = x(.5).$$

The hypothesis that $x_F = .75$ has, similarly, a likelihood density of $x(.25)$. The likelihood ratio, used to calculate support for continuous cases, equals $.5/.25 = 2$ (the random variable x cancelling out). Nonetheless, x_F is uniformly distributed on [0, 1], prior to seeing x. By insisting on *experimental densities* for calculating initial likelihood support for hypotheses about the pivotal (x_F in the case above), Hacking links up the Law of Likelihood, the LPI principle and Lindley's results: all in one swoop.

The LPI is defined with respect to a partition, A. As noted, the partition makes use of the random variable to be measured. It builds in the experimental densities. Therefore, using LPI, the support for the pivotal is just that obtained by using the Law of Likelihood. Finally, Hacking's condition that both x and θ cancel out of the likelihood ratio for the pivotal leads to solutions identical to those sanctioned by Lindley. Hence, using LPI yields fiducial probabilities which satisfy Lindley's coherence conditions.

It is important to realize that Hacking's reconstruction does not avoid the one parameter paradox, discussed on p. 130. That this is so is not at all surprising if the Bayesian interpretation is adopted. Then, one sees each of Hacking's and my reconstruction aiming at a list of necessary and sufficient conditions, *formulated in terms of the statistical distributions only*, for adopting an ignorance prior probability. From Lindley's analysis, we see this is to be a uniform, improper prior. That is, the reconstructions work for a *consistent* formulation of the ancient principle of insufficient reason, without appeal to prior probabilities explicitly. As with more naive solutions, the consistency covers only those cases which avoid arbitrary transformations of the parameter under investigation. For the normal distribution, this extends only so far as linear transformations. The aforementioned paradox makes use of a cubic transformation, e.g. from $\mu \to \mu^3$.

Also, by collapsing the several smooth invertibility conditions into the single 'partition' requirement, Hacking misses the extension to the use of exhaustive statistics. Finally, Hacking's dismissal of the multiparameter paradoxes neglects the epistemologically significant 'step-by-step' maneuver. Hacking says,

> When there are two or more parameters and correspondingly many statistics the matter is notoriously intractable. Contradictions have been discovered in the simplest case of the Normal distribution [Dempster]. The Bivariate Normal is a graveyard. A catalogue of corpses has recently been published [Brillinger]. But in all these cases, whether of

one or more parameters, the trouble can be traced to Fisher's over-generous assumption of irrelevance. Every time a contradiction has appeared, some assumption has been made which is stronger than our principle of irrelevance.[11]

This is a rather strong disclaimer of responsibility for all the known difficulties. In fact, the paradox for one parameter fiducial inference (Section 4.8) does not appeal to any new principles. Since Hacking makes convenient use of Lindley's conclusions to assure us of the consistency of his reconstructed fiducial argument, it would appear that he also accepts CC and Bayes' theorem (so long as all probabilities are derived by inferences legitimized through his theory).[12] But CC and Bayes' theorem are all that is required to construct the one parameter paradox. The pivotals used on p. 131 are valid in Hacking's system. Thus, not *all* difficulties with the fiducial argument stem from "some assumption stronger than our principle of irrelevance."[13]

6.2. HACKING'S PRINCIPLE OF IRRELEVANCE AND GENERALIZED INVERSE INFERENCE

Using an insight due to I. Levi, one can typify Hacking's reconstruction of the fiducial argument by noting that he requires a *demonstration* of irrelevance for the detachment of the fiducial step, in order to arrive at a fiducial probability. This is made most clear when one uses Hacking's reconstruction to avoid fiducial inference in cases where Lindley's conditions fail, e.g. Lindley's example of p. 124. The LPI blocks the irrelevance claim by denying the *initial* likelihood for the pivotal variable based on the experimental density. That is, the likelihood ratio, using the experimental density, *cannot* be expressed as a function of the pivotal variable alone. The unknown parameter interferes so that the hypothesis about the pivotal becomes composite (disjunctive) and the Frequency Principle$_2$ cannot be invoked.

Compare this outlook on relevance with that suggested by Hacking for a modified problem: using *random* samples as 'typical' for a simple inverse inference, from sample to population. A sample is *random* (for Hacking) because of its genesis. He says,

If the chance of singling out any one individual on any trial equals the chance of singling out any other, such a device will be called *a chance set-up for random sampling with replacement*. More briefly, call it a set-up for random sampling, or a random sampling device. A sequence of results from consecutive trials on such a device will be called a *random sample* of the population in question.[14]

A sample is 'typical' when it reflects the parent population. More accurately, a sample is typical with respect to a certain property E when the sample

proportion of E's equals the parent population proportion of E's. This notion can be easily generalized to 'typical $\pm \epsilon$' in the obvious way.

Ceteris paribus, if a sample is *random* then the best likelihood supported hypothesis about its representativeness is that it is *typical*. This follows directly from the definition of *random*. That is, if a closed population of n members has k members with property E, then on a single trial of the random device the chance of getting a member with property E is just $k/n = \theta$. On repeated trials (with replacement) the hypothesis which receives the greatest likelihood support is that the proportion of E's in the random sample is $\hat{\theta} \pm \epsilon$. This is the direct inference, from knowledge of the true parent population to speculation about the random sample. Conversely, for inverse inference, the best likelihood supported hypothesis for θ is that it equals $\bar{x}$, the sample proportion of E's.

This argument is vitiated when the sample is *misleading*. A sample is *misleading* (for Hacking) whenever the hypothesis of typicalness is *not* the best likelihood supported one. This can occur if some special feature of the random sample is discovered. The special feature must be such that it is *known* that the proportion of the parent population with the special feature having E differs from the proportion of E's at large. That is, it is known that the sample has an additional property besides being random. By the Total Relevant Evidence requirement this extra information blocks the argument for typicalness, given above.

The epistemological condition is required since, without it, no sample would ever be non-misleading. For example, the property of having been selected by the investigator is not common to the members of the parent population, at large. But it is not known that merely being selected by the investigator alters the proportion of E's so produced. Hacking puts the point this way,

> This bears on another matter. Suppose an experimenter completes his experiment, and infers from the sample that the best supported hypothesis about the proportion of E's is that it is p. Someone else notices that every member of his sample has some property B which no one would be inclined to account a 'trick' property like the one I just described [being selected by the investigator]. There is still no grounds for criticism of the experimental result. It is of course open to the sceptic to try another sample in which not all members are B's, but unless there is good reason to suppose, for instance, that $P(E)$ differs from $P(E/B)$, it remains the case that on the data the best supported hypothesis about the proportion of E's is p.[15]

Now this seems like a perfectly reasonable explanation of why most extra information about random samples fails to prevent them from being considered typical. More accurately, the extra information does not make them

(automatically) misleading. This allows Hacking to claim that the two ways for criticizing inference from sample to population are:

> (1) the sample is not random; (2) available data show the sample is misleading. You can go wrong either through dependent trials, or by the bad luck of drawing a misleading sample.[16]

The shift in the analysis of relevance is clear. In the fiducial case, all irrelevance has to be demonstrated. In the inference from sample to population, the charge of misleadingness has to be demonstrated. In the fiducial case, if the unknown parameter interferes with the likelihood, then irrelevance is not demonstrated, hence, not granted. In the random sampling case, if the sample shows some special property B and the probability $\mathrm{p}(E/B)$ is not known, misleadingness is not demonstrated; hence B is not demonstrated *relevant* for inference from sample to population.

Both fiducial probability and inference from sample to population are instances of inverse inference. The preceding analysis shows that Hacking cannot claim a unified theory of relevance for inverse inference. However, he tries to put the two together.

The first instance of a 'fiducial like' argument Hacking presents is the following, heuristic example:

> Imagine a box and an urn. The box contains a coloured ball. You want to know its colour but cannot open the box. You are told that the urn contains 100 balls; 95 of the balls are the same colour as that in the box. The chance of drawing a ball of some colour is exactly in proportion to the number of balls of that colour in the urn. Hence the chance of drawing a ball the same colour as the boxed ball must be .95. Suppose one ball is drawn, and it is red. Everyone agrees it is a good bet, that the boxed ball is also red. Indeed .95 seems to measure the degree to which the whole set of data supports the hypothesis, that the ball in the box is red. Can this be justified?[17]

Hacking goes on to duplicate a fiducial argument setting up a pivotal-like proposition H: The ball drawn from the urn is the same colour as the ball in the box. Then $\mathrm{p}(H/BK) = .95$. Let e be: The ball drawn is red. By an irrelevance claim, $\mathrm{p}(H/BK \,\&\, e) = .95$ too. But, given $BK \,\&\, e$, H is equivalent to the hypothesis that the boxed ball is red. Hence, a posterior probability for that hypothesis, in the absence of a prior: all in the spirit of the fiducial argument.

Suppose, however (as is reasonable) the sampled ball is seen to have a scratch mark on it. That is, the total evidence is e & 'the sampled ball is scratched'. What of the fiducial argument now? Using the recommendation appropriate for inference from sample to population, the extra evidence is not demonstrably relevant, e.g. it is not known what proportion of scratched

balls, sampled from the urn by the device, are of the same colour as the boxed ball. But, by the analysis from the continuous fiducial argument, neither is it demonstrably irrelevant. Therefore, given that the new pivotal is modified by the information of the scratch, the initial support is undefined, and the irrelevance claim *not* granted. Which way does Hacking want it? It can't be both.

As remarked before, the change in onus reflects the desire to be able to disregard extra information when worried about 'typicalness.' The information suppressed is not known to be relevant, hence no outright violation of TRE is made. On the other hand, the coherence of the fiducial argument (based on Lindley's work) shows that a strict accounting of irrelevance is needed. In fact, the questionable cases, i.e. where no demonstration is possible either way, must be counted *against* the irrelevance claim. Either prove the irrelevance, or else deny the fiducial argument its irrelevance claim, needed to detach in the Fiducial step.

Thus, Hacking has salvaged the fiducial argument only at the cost of a unified theory for inverse inference. That this dichotomy is necessary, i.e. that random sampling and fiducial conclusions cannot be defended with a unified theory of relevance, will be made clearer after the next chapter in which Henry Kyburg's program is discussed. Unlike Hacking, he tries to cover both topics with a theory of relevance such that the onus is on the claim of relevance. That is, all relevance claims must be demonstrated. As would be expected from the hints made here, subject to the CC principle, the program falters when the coherence problem is faced.

NOTES

[1] Ian Hacking, *Logic of Statistical Inference* (Cambridge: Cambridge University Press, 1965).
[2] *Ibid.*, p. 59.
[3] *Ibid.*, p. 135.
[4] *Ibid.*
[5] See Appendix 9.6.
[6] *Ibid.*, p. 145.
[7] *Ibid.*
[8] Hereafter it is assumed that, for continuous cases, partitions need not be *countable*.
[9] However, density ratios *are* invariant to (differentiable) *calculable* 1-1 transformations! See Appendix 9.4 for a proof that Hacking's fiducial argument is invariant for these.
[10] Hacking, *op. cit.*, p. 69.
[11] *Ibid.*, p. 151.
[12] *Ibid.*, pp. 154–155.

[13] Hacking's remarks about multiparameter fiducial arguments are terribly obscure. He says little, if anything at all, about the step-by-step method. Matters become even more confused by an allegedly valid joint fiducial argument for the Normal distribution, *ibid.*, pp. 156–158; even when Hacking apparently recognizes the seriousness of Dempster's alternative (incompatible) solution (see the quote, above). How does Hacking choose to remove the conflict with Dempster's results? No answer is found in the text.

Another question for Hacking's reconstruction, which bypasses the step-by-step method, is how to handle Tukey's paradox (of page 156)? The solution offered here (page 157) depends on the avoidance of incalculable transformations of the *r.v.* in conjunction with the step-by-step technique. It is not clear (due to Hacking's casual treatment of the multiparameter version of his own argument) whether or not the experimental densities for Tukey's (transformed) pivotal are sensitive to these issues.

The upshot of these remarks is that Hacking's reconstruction is deficient for the multiparameter extension of the fiducial argument. It is, of course, open to Hacking to adopt the suggestions proposed here in clarification of the more complicated problems.

[14] *Ibid.*, p. 122.

[15] *Ibid.*, p. 128.

[16] *Ibid.*

[17] *Ibid.*, pp. 136–7.

CHAPTER 7

HENRY KYBURG'S THEORY

Henry Kyburg's program of inductive logic is unique for its blend of appealing philosophical ingredients. All probability statements are of a *direct probability* form, similar to Fisher's later views (see 3.1). Probability is *Frequentist* in style, since relevant evidence is summarized as (intervals) of class ratios. No *stochastic* (probability$_2$) concept is required, as class ratios alone lead to probability$_1$ relations through the notion of *randomness*. Also, since it is unreasonable to expect knowledge of class ratios to be exact all the time, Kyburg uses interval-valued randomness relations to support *interval-valued probability*.

The most surprising consequence of Kyburg's position for fiducial theory (as discussed in 7.1) is that fiducial probability requires no new principles of relevance (unlike Hacking's formulation). All the relevance principles are built into the randomness constraints, which pick out the valid direct inferences. Whereas Hacking's Irrelevance Principle sets conditions for demonstrating irrelevance, i.e. irrelevance must be shown, the onus is reversed in Kyburg's rules for randomness, i.e. relevance must be shown. Hence, where no clear demonstration is possible, Hacking treats the data as relevant to the pivotal variable and Kyburg treats them as irrelevant.

I. Levi has shown that this approach to direct inference leads Kyburg's theory into a violation of conditionalization. In the text, I show that the violation is so pervasive that even the Sufficiency Principle is invalid for Kyburg's system. (This is a somewhat stronger result than Levi's, as conditionalization entails sufficiency, but the converse does not hold.) Thus Kyburg's reconstruction is found to be faulty in excess of the deficiency common to Hacking's and my own attempted reformulations. Of course, the difficulty encountered in 4.8 also affects Kyburg's version of fiducial inference.

Although the analysis of 7.1 reveals the inadequacy in Kyburg's treatment of fiducial inference, it suggests that his position might be amenable to randomization. In 7.2 I consider Kyburg's views on randomization, and show that the relevance principles which caused the faults in rebuilding fiducial theory are suited for a limited defense of randomization in design.

7.1. HENRY KYBURG'S 'RATIONALLY REPRESENTATIVE' PREDICATE AND THE FIDUCIAL ARGUMENT

Henry Kyburg's program for a theory of statistical inference, although frequentist in its outlook, has several very novel aspects which distinguish it from any of those considered previously.[1] For the discussion of the fiducial argument, perhaps the most important feature (first pointed out to me by Isaac Levi) is its attitude towards relevance. That is, (unlike Hacking's) Kyburg's position requires that relevance must be demonstrated, and in the absence of such a demonstration irrelevance is assumed.

CHAPTER 7

Before reviewing Kyburg's reconstruction of fiducial inference, it seems to me best, first to summarize several of the more significant individuating qualities of his approach. The goal of this critique is to clarify the program's position on the two issues in inverse inference developed in the last chapter, namely, the fiducial argument and inference from sample to population (for 'typicalness') in general. The conclusion to be defended is that, although the explication of relevance is uniform over both problems, it is satisfactory only with respect to the second. The fiducial argument (under Kyburg's reconstruction) violates conditionalization. This violation is traced, with an appeal to analysis due to Isaac Levi, to a general failure of CC strong enough to void the claims that sufficiency and ancillarity are legitimate relevance and irrelevance conditions. The source of the failure of CC in Kyburg's theory is his solution to direct inference.[2]

The first point of departure from the usual treatment is the absence of stochastic (probability$_2$) statements. Hacking's Frequency principle, adopted for my reconstruction of Fisher's views on direct inference (Section 3.1), makes explicit reference to a chance or stochastic property as part of the evidence. Kyburg admits a confirmational interpretation of probability, but the data he uses for direct inference are simple frequencies without any probability$_2$ connotation. For instance, if marble M is known to be one of four, differently colored marbles, if one is known to be red, then (on Kyburg's view) the probability is $\frac{1}{4}$ that M is red. Specifically, it is not necessary for valid inference that M be selected by any *process* we would call 'random'.

This leads to a very interesting generalization of probability. Instead of demanding that the evidence consist of precisely known frequencies, Kyburg permits the use of interval valued frequencies to generate interval valued probabilities. The advantage of this approach is that epistemic situations too weak to support precise probabilities now are represented by interval valued probabilities. Since the frequencies do *not* have to be backed-up by the extra burden of a stochastic assumption, it follows that Kyburg's interval probabilities can be defined on any background knowledge, not just those with exact statistical evidence.

In those cases, the classic ones, where statistical knowledge is presupposed, the corresponding Kyburgian evidence is merely the frequency part of the stochastic information. For example, to say that this is a fair coin when flipped by device F (a stochastic, binomial claim) is replaced in Kyburg's system by the claim that the appropriate reference set for direct inferences about flips of this coin consists of 50% 'heads' and 50% 'tails'. Now, as is well known, it is possible to flip a fair coin any number of times without getting

an observed frequency of exactly half 'heads'. But in Kyburg's program the reference sets must refer to specific trials and, for a fair coin, Kyburg presumes that his regulations pick out a reference set with just the right frequencies. These regulations are decidedly epistemic, paralleling Reichenbach's 'narrowest reference class with known statistics' or Fisher's 'recognizability'. A more detailed examination of Kyburg's regulations for selecting a reference set follows shortly. However, it is an interesting and important point to show that even knowledge that this fair coin is to be (or has been) flipped just once does not automatically interfere with the usual direct inference conclusions, and for this a rigorous understanding of Kyburg's rules is not needed.

The focal point of Kyburg's program, on this issue, is the denial of any stochastic or chance mechanism peculiar to this coin. That is, to repeat, it is only the frequencies within reference sets (or, more accurately, what is *known* of these frequencies) that counts. Thus, one is not limited to the trivial reference set of the unit set of the only flip of this fair coin. One may consider the more extensive reference set of flips with coins *similar* to this one. Of course, it is the epistemological factor which determines the bounds for this similarity. Typically, i.e. with usual background information, the reference set for the problematic case (above) would consist of the outcomes of flips of ordinary coins, generally, and it is presupposed that it is *known* that very nearly 50% of these land 'heads-up'.

A second significant part of the Kyburgian system is the careful use of a formal language to represent all statements, both probability and non-probability ones. In this sense the meaning of the *logical* status for probability is more clearly presented. Although the implications of this restriction to a formal language are not at once obvious, it does play a major role in preventing the collapse to degenerate [0, 1] probabilities. In lieu of a rigorous analysis of this feature, we can note that its primary importance for the discussion here is to bar certain reference sets by eliminating the corresponding reference terms from the language.

In traditional frequentist spirit, the biggest hurdle in the construction of Kyburg's probability statements is the determination of the reference set. In untraditional fashion, the conflict is not among competing probability statements, but instead among competing randomness claims; which are the precursors of probability. Thus, all the usual concerns find their way into the explication of randomness.

For Kyburg, randomness is a six termed relation. Given a formal language L, and with respect to a background knowledge w (assumed consistent), an entity x is known to be a member of the set y whose **Z**-value is claimed to

belong to the set (of values) b. In other words, $\mathbf{Z}(x)$ is a random variable for the entity x, and b is a subset of the sample space for z. The randomness statement, denoted $\mathbf{RAN}_L(x, y, z, b, w)$, asserts that this membership of x in y is *random* with respect to z and b.

The randomness statement is backed-up by a statistical one (contained in w), which gives the frequency range for the statement: $\mathbf{Z}(x) \in b$, for the reference set y. It is denotes $\mathbf{S}\langle y, z, b, p_l, p_u\rangle$; where p_l is the lower frequency and p_u the upper frequency. In addition, it is supposed that there is a uniquely strongest statistical statement s about y, z and b (denoted $\mathbf{PS}_{w, L}(s, y, z, b)$) having frequencies $[p_l, p_u]$, and these agree exactly with $\mathbf{S}$ (above). That the statistical statement is *strongest* means that no other statistical statement about y, z and b is known to have frequencies $[p'_l, p'_u]$, where $p_l \leqslant p'_l$ and $p'_u \leqslant p_u$ (with at least one inequality strict). That the strongest statement is *unique* means that no other statistical statement about y, z and b conflicts by having a *different* interval $[p''_l, p''_u]$; where *difference* occurs so long as one interval is not stronger than the other and they are not identical. In other words, either the two intervals are disjoint or else they overlap without containment. To repeat, that there is a uniquely strongest statistical statement presupposes no conflict due to difference or strength.[3]

If, given the context, a particular randomness statement is unchallenged or undefeated, the derivative probability statement is just the natural one. That is, the probability that x has a $\mathbf{Z}$ value in b, relative to w is $[p_l, p_u]$. The frequencies are those of the statistical statement used to support the randomness claim. Comparing this to the 'definition' of probability credited to Fisher (see pp. 72–4), the recognizability condition for the reference set is replaced by a randomness requirement in Kyburg's definition.[4] However, unlike Fisher's rather vague notion of recognizability, Kyburg is quite explicit in his analysis of randomness.

One key feature is common to both ideas and, ultimately, is the source of the title 'epistemological probability.' Randomness, like recognizability, is an epistemological condition. More precisely, the question of which statement is to serve ultimately as the randomness one, and hence which set y (known to contain x) is to serve as the reference set, is a function of w alone, i.e. what is known. The overriding principle for determining the reference set is to allow only that randomness statement which generates the most precise statistical statement (subject to the extensiveness of the reference set).

For example, if a card has been dealt from a standard deck after shuffling, then the usual probability assignment to the hypothesis that the card is the ace of spades is 1/52. Working backwards in Kyburg's program, the usual

randomness statement for this probability must be one which assigns a frequency of 1/52 to some reference set (yet to be determined) with respect to being an ace of spades and such that this card, just dealt, is known to be a member of it. A natural candidate is the set of cards dealt on repeated trials, under the assumption that exactly 1/52nd of these will be aces of spades. This last assumption coincides with the belief that the dealing process is a 'fair' one, i.e. one that selects each card equally often on repeated trials. However, if the usual background knowledge is modified, so that it becomes known that the card dealt is an ace, then the narrower reference set, consisting of just aces dealt, is the appropriate one. Hence, with the enlarged background information, the probability is $\frac{1}{4}$ that the card is the space ace.

All this is not surprising and without conflicting desiderata. The extra information is used to narrow down the reference set, and the resulting frequencies are just the conditional frequencies within a sub-reference set of the original (larger) one. But it must be remembered that Kyburg's program has a second dimension for representing knowledge. The frequencies within a reference set may be known only imprecisely, i.e. by the bounds $[p_l, p_u]$. Assuming the language to be even moderately rich in reference terms, it is a virtual certainty that some narrower reference set than the usual one can be found. At the very worst, the knowledge of frequencies for this new reference set is [0, 1], the logically true range for any set. Since the reference set need not carry a stochastic property; that is, the frequencies are all that are needed for evaluating the randomness statement, the additional interval valued randomness statement-candidates makes identification of the overriding 'narrowest reference set' less obvious.

For example, the trivial unit set of the event in question is readily available as a potential reference class, with references known as [0, 1]. In the card example, the set having only one member, the card dealt, may well serve as a reference set. Presumably, its frequencies are known (from w) only to be [0, 1]. Thus, the desire for a precise statistical claim runs afoul of the demand for a narrow reference set (which can reflect relevant knowledge through conditionalization – but remember that Kyburg does not support CC generally.)

Kyburg resolves the dilemma in favor of the precision of the statistical claim. Thus, his *prevention by strength* rule decides between competing randomness claims which trade off narrowness of the reference set (through the subset relation) against imprecision of the statistical knowledge (through the size of the interval expressing the frequencies).

Principle I (prevention by difference): If a is known to belong to both X and Y, and our statistical knowledge about X and Z differs from that about Y and Z – i.e., if one interval

is not included in the other – then, unless we happen to know that X is included in Y, Y prevents X from being an appropriate reference class for '$a \in Z$'. If neither is known to be included in the other, then each prevents the other from being an appropriate reference class.

Principle II (prevention by strength): Given that we want to avoid conflict of the first sort, we want to use the most precise information we have; and therefore among all the classes not ruled out by each other in accordance with Principle I, we take as our reference class that class about which we have the most precise information.[5]

The first rule captures the intuition that more knowledge, reflected in a narrower reference set with conflicting statistics, takes precedence. The second rule, less intuitive but equally essential to get Kyburg's system off the ground, ranks narrower frequencies over broader, containing frequencies (among sets not eliminated by Principle I), and therefore rules in favor of randomness claims with point valued statistics which survive difference considerations (Principle I) over those relying on an interval containing the point. Hence, usually a reference class with less than trivial, [0, 1], frequencies wins out over one with the degenerate, logically true range [0, 1]. It follows that unless the unit set has a known exact frequency, e.g. the outcome is either known to be the ace of spades ($p_l = 1$) or known to be a non ace of spaces ($p_u = 0$) the unit set will fail to some other by the strength rule.[6]

This sketchy review of Kyburg's program is already sufficient for reconstructing the fiducial argument. However, in preparation for the general analysis to follow, it is appropriate first to consider a restriction relevant to statistical inference. As we remarked earlier, Kyburg does not require a probability$_2$ notion so that finite reference classes and finite frequencies are admissible. For example, he says,

If I am told that ten tosses of a certain coin were performed yesterday and their outcomes recorded, and that seven of the ten resulted in heads, then 0.7/0.3 represents the appropriate odds for me to offer on tails on the third of these tosses, even if I know for sure that the coin and the tossing apparatus are perfectly fair. But *in general* when confronted with gambling apparatus or a physical set up embodying a statistical law of physics, it is true that with respect to a given trial, at a given time, one of two epistemological states obtains: either we know the outcome of that trial, or there is nothing we know that keeps that trial from being a random member of the general class of such trials.[7]

It is agreed that ignorance about the bias of a strange coin is to be captured by the vacuously true [0, 1] frequencies, assigned to the reference set of, say, flips with the coin. Another model for the binomial distribution is sampling from an urn with a finite population and with fixed frequencies for

some property of the chips contained therein. Stepping out of Kyburg's theory and reverting to our use of stochastic properties, we can say that just as the chance of heads is a property (dispositional, perhaps) of the coin and flipping device, so too the stochastic binomial chance of getting a chip of a certain kind from the urn is a composition property of the proportions within the urn and the sampling device for obtaining chips from the urn.

Ignorance about the chances may reflect ignorance about the sampling device alone. The proportion of chips with the property in question may be known. To insure that the randomness statement about the next chip to be drawn includes only the [0, 1] frequencies, the population of chips must be barred as a reference set; else the proportion within the population would be available as a frequency which could win out by 'strength'. This would induce a sharp, precise representation for ignorance, against the doctrines of the program.

Kyburg responds to this charge by saying,

> It is clear that far from requiring an equal distribution of ignorance regarding the possibilities, this mode of analysis requires – not an equal distribution of knowledge – but knowledge of an equal distribution of frequencies among the possibilities. Can we be quite sure that something amounting to the bare counting of possibilities won't arise under the interpretation of probability that has been offered? For example, one could consider the increasing sequence of sets of rolls with a die as the increasing sequence of sets of numbers yielded by the die. (Or, the increasing sequence of values of the random quantity representing the number of spots on the die.) The union of the range of this sequence is just the finite set $\{1, 2, 3, 4, 5, 6\}$. This is a mathematical object, and '$\{1, 2, 3, 4, 5, 6\}$' is a mathematical term. But scrutiny of MD-7.39 and MD-7.40 will reveal that although mathematical terms may enter intersectively into reference terms (as in the set of all n-fold subsets of S) they do not by themselves qualify as reference terms. Thus we cannot have as a reference term, a term which denotes simply a set of 'possible outcomes', unless we can construct such a term from the primitive terms we have available. And there seems to be no reason to use a language in which this is possible.[8]

Therefore, for statistical distributions, arbitrary reference terms, e.g. the sample space, do not automatically qualify as legitimate reference sets. Of course, probability statements using an admissible reference class may be abbreviated by talk about the probability for outcomes in the sample space. But this is not to be misconstrued as the use of the sample space as a reference class.

Now for the reconstruction of the fiducial argument within Kyburg's theory. Instead of beginning with a continuous case, true to Fisher's dictum, rehearse Kyberg's argument for a discrete binomial example, thereby making use of the extra interval-valued machinery.[9] The focus for Kyburg's

reconstruction is the 'Rationally Representative' predicate, which takes the place of the pivotal variable.

Suppose that a strange coin has been flipped by device F 100 times with an observed frequency of 60 heads and 40 tails. The background knowledge w contains only the binomial assumption. That is, the prior probability statements about the bias of the coin have logically vacuous [0, 1] intervals. Define a property of a binomial sample, called the rationally representative property, as follows:

Let $\bar{x}$ equal the sample proportion of heads, and θ the unknown bias for heads using the F device. A sample is ϵ-rationally representative if $\bar{x}$ differs by no more than ϵ from θ, i.e.

$$|\bar{x} - \theta| \leqslant \epsilon. \qquad \text{[RR]}$$

Now, before seeing the actual sample, an interval valued probability statement about its RR-ness is available. The argument goes like this: For moderately large samples from a binomial population, the sample average is normally distributed $N(\theta, 1/[\theta \cdot [1-\theta] \cdot n])$, that is, with mean θ and variance $1/(n\theta - n\theta^2)$, where n is the sample size. The proportion of samples (of 100) which are ϵ-RR is known once θ is fixed. This propertion reaches a minimum for $\theta = .5$; whence the variance is a maximum, $4/n$. Therefore, if this proportion is $1 - \delta_\epsilon$, it is *apriori* true that the proportion of ϵ-RR samples (of size 100) is no less than $1 - \delta_\epsilon$, or $[(1 - \delta_\epsilon), 1]$.

Assuming that w contains no additional information about θ, then note that prior to knowing the sample outcome the proportions $[(1 - \delta_\epsilon), 1]$ can serve as the frequencies for a randomness statement about the ϵ-RR-ness of the sample to be selected. The final step in the reconstruction is to consider this statement after observing the sample and argue that the randomness claim is not prevented even after $\bar{x}$ becomes known.

First, consider what is the consequence of retaining the claim of RR-ness of the sample, after observing $\bar{x}$. Once $\bar{x}$ is known, it becomes a constant so that the probability is at least $(1 - \delta_\epsilon)$, i.e. $[(1 - \delta_\epsilon), 1]$, that $(\bar{x} - \epsilon) \leqslant \theta \leqslant (\bar{x} + \epsilon)$. But this is a fiducial-like posterior probability for θ which is relatively precise, i.e. $1 - \delta_\epsilon$ is much greater than 0, based on an imprecise 'ignorance' [0, 1] prior probability.

How can the rationally representative status of the sample be sustained *after* observing the sample proportion $\bar{x}$? The answer depends on the Kyburgian approach to relevance. Since randomness, unchallenged, converts into probability, it requires a competing randomness claim which prevents the RR-ness of the sample to deny the fiducial-like probability. In

the absence of a successful challenger, the original randomness statement survives and the desired conclusion is detached.

Kyburg's system employs numerous prevention and counterprevention rules for determining the validity of any randomness statement. Between the total absence of knowledge of $\bar{x}$ (the epistemic state prior to seeing the sample) and precise knowledge of $\bar{x}$ (the epistemic state after seeing the sample) there is a large family of knowledge states about the sample proportion. Each of these intermediate states can generate a competitor randomness statement to the RR claim. That is, by utilizing less than full knowledge of $\bar{x}$ a potentially relevant reference set for statements about θ is obtained. It is Kyburg's claim that none of these prevents randomness, using the RR predicate, after $\bar{x}$ is observed.[10]

To defend Kyburg's claim, and also to emphasize the extreme importance of the prevention by strength rule for the argument, we will consider one case. Imagine the sample taken and $\bar{x}$ made known. At one end of the spectrum of possible challenger randomness statements is the one based on complete knowledge of $\bar{x}$. That is, in place of RR-ness, consider the potential randomness statement based on the proportion of 100 fold samples, *all of whose heads/tails proportion is* $\bar{x}$, which are ϵ-RR. That is, use the extra knowledge of $\bar{x}$ to restrict the reference set of samples to those with a sample proportion of exactly $\bar{x}$.

In this new reference set the proportion of ϵ-RR samples is modified. If $|\theta - \bar{x}| \leqslant \epsilon$, i.e. for populations with a θ value within ϵ of $\bar{x}$ (now, a constant), then *all* such samples are rationally representative. However, if $|\theta - \bar{x}| > \epsilon$, i.e. for populations with a θ value more than ϵ from $\bar{x}$, then *no* such sample is rationally representative. Therefore, the overall proportions of ϵ-RR-ness for the restricted reference class are $[0, 1]$, the vacuous ones.

Looking back at the prevention by difference rule (p. 188), we see that the restricted reference set does not prevent by difference, since $[(1 - \delta_\epsilon), 1]$ is contained within $[0, 1]$. However, the prevention by strength rule *does* pick out the unrestricted reference class as the correct one for the correct randomness statement. Therefore, knowledge of $\bar{x}$ does not prevent the randomness of the rationally representative statement and the fiducial-like probability is valid. In other words, $\bar{x}$ is not demonstrated to be relevant to θ (through the RR predicate), hence is irrelevant.

Kyburg's reconstruction for exact fiducial probability arguments depends, of course, on the acceptance of continuous distributions. Although this acceptance is made with only minor reservations by Kyburg,[11] it does seem to run against the grain of the finite frequency approach. Taking Kyburg at

his word, however, for continuous distributions the interval frequencies $[(1-\delta_\epsilon), 1]$ of the RR predicate are replaced with exact ones. For example, if x (the *r.v.*) is known to be normally distributed $N(\mu, 1)$, with only degenerate $[0, 1]$ frequency information about μ, then the fiducial argument is obtained from the RR predicate

$$|\mu - x| \leqslant \epsilon. \qquad \text{[RR]}$$

Unlike the binomial distribution, the frequency distribution for *normal* representativeness has constant (unit) variance, independent of the unknown mean μ. Therefore, the proportion of ϵ-RR samples (of 1) for the normal distribution is given precisely, i.e. $[\delta_\epsilon, \delta_\epsilon]$, once ϵ is set. The argument for randomness, after seeing x, is identical to the binomial case, only with precise point-valued frequencies in place of the interval ones.

To repeat, the RR predicate takes the place of the pivotal. The restrictions on the pivotal from the previous reconstructions, e.g. 'naturalness' for ours from chapter 4, and through the LPI for Hacking's, are duplicated in the relevance conditions fomulated in terms of the randomness requirements. Unlike Hacking's program, the fiducial step now is accomplished by *failing* to show the relevance of the extra information available once the sample is observed.

This change of onus from Hacking's reconstruction, although compatible with a program of inference from sample to population (as demonstrated in the next section), is suspicious. The fiducial argument is duplicated within Kyburg's theory without attention to the various threats of incoherence, e.g. Lindley's analysis. Part of the load is carried by the strength rule and it is doubtful that this one, relatively novel condition can replace the more complex ones used in the other reconstructions. In fact, I. Levi has recently shown that the strength rule (in conjunction with the difference principle – which by itself is a plausible relevance condition) leads to inconsistencies for direct inference if CC is accepted. Of course Kyburg chooses to reject CC, but Levi's analysis shows that there can be no choice in the matter if Kyburg retains his relevance conditions, i.e. his randomness principles.[12]

In the following brief, but technical, discussion I will try and argue that this failure of CC is severe enough to void the claim that Sufficiency is a valid relevance condition. That is, the strength rule is so pervasive that Kyburg's system does not select arguments based on sufficient statistics over different arguments using less than sufficient statistics. This result should not come as a surprise if it is remembered that our justification for Sufficiency rested on CC (p. 20) and that Levi's argument shows that CC is violated even in simple cases.

Our example involves Kyburg's version of fiducial inference and proceeds by considering two different RR predicates for the same (unknown) parameter. Only one of these predicates uses a sufficient statistic for the sample data. I argue that Kyburg's randomness conditions are incapable of picking out the argument based on the sufficient statistic as the only legitimate randomness statement. Hence, two incompatible fiducial conclusions threaten. Coherence is saved within his system by denying both randomness claims and reverting to degenerate [0, 1] probability statements.

The problem is to obtain probability statements about an unknown mean μ, using a sample of n independent readings (n odd) from a normal distribution with unit variance, i.e. $N(\mu, 1)$. Ignorance of the mean is captured in Kyburg's system with the vacuous [0, 1] frequencies. Consider two reductions of the data: $\bar{x}$ the sample average (which is sufficient) and m the sample median (which is not sufficient).

Each of the statistics can be used to construct a pivotal variable (involving μ). Define the two RR predicates as follows:

$$|\bar{x} - \mu| \leqslant \epsilon \qquad [RR_1]$$

$$|m - \mu| \leqslant \epsilon. \qquad [RR_2]$$

RR_1 uses the usual pivotal for this problem (see the example on p. 108) which is distributed $N(0, 1/n)$. The median pivotal for RR_2 is also normally distributed $N(0, (\pi/2) \cdot (1/n))$, but with the greater variance. Technically, the median is a consistent and *not* efficient statistic for μ.[13]

Note that both RR statements are based on the same reference class, i.e. samples of n from the normal distribution $N(\mu, 1)$, and $\bar{x}$ and m are calculated from the same sample data. Also, both pivotal distributions are precise so that the strength rule does not operate (directly) to select one RR randomness statement over the other. However, the fiducial probabilities induced for μ by these two pivotals are never in agreement. Thus, unless at least one randomness claim is denied incoherence follows.

For Kyburg's theory, two claims must be defended if RR_1 randomness is retained. First, RR_1 statements must be declared random even after m becomes known. Second, RR_2 statements must be declared non-random once $\bar{x}$ becomes known. No such problems arise for the other reconstructions as the Sufficiency principle leads to the rejection of RR_2.

The first of the two claims: that knowledge of m does *not* interfere with statements of RR_1 randomness, will not be examined here. At the time of this writing and after correspondence with Kyburg no argument in support of

this claim is available. The difficulty involves another of his randomness conditions (the cross product rule), which has not been discussed in this work. However, the second claim proves more interesting as the problems which arise when trying to defend it result from the strength rule.

That $\bar{x}$ is sufficient for μ, while m is not, can be seen by considering the joint distribution for $(\bar{x}, m)$:

$$\mathrm{p}(\bar{x}, m/\mu) = \mathrm{p}(m/\bar{x}) \cdot \mathrm{p}(\bar{x}/\mu) \qquad [*]$$

$$= \mathrm{p}(\bar{x}/m, \mu) \cdot \mathrm{p}(m/\mu). \qquad [**]$$

Note that in [*] μ appears in only one factor, but that in [**] it appears in both. The RR_1 pivotal is based on $\mathrm{p}(\bar{x}/\mu)$ (of [*]) and RR_2 is based on $\mathrm{p}(m/\mu)$ (of [**]).

In order that the second claim be defended within Kyburg's theory, a new reference set must be provided such that it defeats the RR_2 randomness statement. However, to argue that no such reference exists is to make a universal statement. Like the example on p. 191, one must consider a family of reference sets based on the knowledge of $\bar{x}$. Instead of examining many cases in the spectrum of reference classes, consider one extreme (corresponding to conditionalization) in which the reference set is selected with $\bar{x}$ fixed at its observed value.

From [*], one sees that if the reference class is restricted by holding $\bar{x}$ constant the resulting (conditional) distribution for m is not $\mathrm{N}(0, (\pi/2) \cdot (1/n))$. But does this change result in a relevant difference, in Kyburg's technical sense of 'difference'? Specifically, the distribution of m, with $\bar{x}$ held fixed, is *independent* of μ (from the sufficiency of $\bar{x}$ for μ). The distribution of m in samples of n (n odd) is $\mathrm{N}(\mu, (\pi/2n))$, whereas that for $\bar{x}$ is $\mathrm{N}(\mu, (1/n))$. Thus the m distribution has the larger variance. From [*] it follows that the variance for m can be represented as the sum of two parts: the variance of $\bar{x}$ from μ, plus the variance of m from $\bar{x}$ (call this $V_{m/\bar{x}}$). That is, $(\pi/2n) = (1/n) + V_{m/\bar{x}}$. So, $V_{m/\bar{x}} = (\pi - 2)/2n$, which is independent of the unknown μ. Therefore, a close approximation for the distribution of m, given $\bar{x}$, is $\mathrm{N}(\bar{x}, [(\pi - 2)/2n)]$. What is the import of this for the randomness claim? If $|\mu - \bar{x}| \gg 0$, i.e. the observed $\bar{x}$ is far from μ, then since the conditional distribution for m keeps it 'close' to $\bar{x}$, the frequency for $|\mu - m| \leqslant \epsilon$ is negligible. In other words, the lower frequency bound for RR_2 restricted to a reference set with constant $\bar{x}$ is very nearly 0.

Now, since the variance for m, given $\bar{x}$, is less than the unconditional variance for m, if the case $\bar{x} = \mu$ is considered, then the conditional distribution for

RR_2 (in the new reference class) is modified so that the frequency for $|\mu - m| \leqslant \epsilon$ is increased. That is, if ϵ is set so that the proportion of RR_2 samples is .95, then when the reference set is conditioned on a constant value for $\bar{x}$ (which now equals μ), the frequency is greater than .95.

Thus, in the restricted reference class the proportions of RR_2 samples are interval valued, and the interval includes the point value from the unconditioned reference class.[14] Citing the strength rule, the narrower reference class (from knowledge of $\bar{x}$) does *not* prevent the randomness statement about RR_2, using the larger reference class. The second claim cannot be defended through appeal to knowledge of $\bar{x}$.

Since $\bar{x}$ is sufficient for μ, while m is not, the CC principle (along with Bayes' theorem) entails the relevance of $\bar{x}$ for μ, even given m. The converse claim, i.e. the irrelevance of m for μ, given $\bar{x}$, is also true (by Sufficiency). That strength interferes with conditionalization, in general, is Levi's conclusion. That it also leads to a failure of Sufficiency is demonstrated by the RR_1-versus-RR_2 example. Without Sufficiency, not only does Lindley's paradox emerge, but also a wholly new collection of problems similar to the mean-median problem. It seems reasonable too, that without an appeal to the CC principle, no justification for the 'step-by-step' procedure is available. One can always supplement the Kyburgian randomness condition with specific recommendations for each case; however, the piecemeal compensation for the failure of CC looks hopeless. Therefore, I believe that Kyburg's reconstruction of fiducial inference is inadequate despite its clear cut frequentist basis.

Of course, Kyburg is not ignorant of the violation of CC that follows from his randomness principles. In fact, he argues that it is an advantage![15] Although I do not agree with him regarding the importance of CC for fiducial inference, I must admit that his program does offer advantages for what we called general inverse inference, from sample to population (pp. 178–181). In the section that follows I examine the upshot of his randomness principles for problems involving random sampling.

7.2. KYBURG'S RELEVANCE CONDITIONS AND RANDOMIZATION

In the previous section I argued that the strength rule was the major factor causing the failure of the Sufficiency principle in fiducial inference within Kyburg's theory. Now we shall see that it is the same rule which leads to an unusual treatment of randomization (in experimental design). In Hacking's

theory we discovered that it was necessary to change the relevance conditions to accommodate both fiducial inference and inference from sample to population for 'typicalness' (pp. 178–191). Specifically, irrelevance had to be demonstrated for fiducial inference, while relevance had to be demonstrated for general inverse inference. Since, in Kyburg's system the latter position is taken without exception (it arises from the strength rule), his program permits the sorts of conclusions advocates of randomization seek. Our central question is, in which situations can randomization support inductive inference which otherwise would be invalid?

Perhaps the most famous and pedagogically valuable example of randomization in design (as a replacement for ignorance) is Fisher's experiment about the lady who claims to taste the difference between tea prepared by adding milk first and tea prepared by adding tea first.[16] In this classic problem of design, Fisher hopes to perfect a significance test on the (null) hypothesis that the lady cannot discriminate tea by the order of preparation.

Fisher constructs his test according to the following prescription. Eight cups are to be presented to the lady, four made one way and four made the other way. She is to be asked to sip each and then divide them into two groups, according to her judgement of which were made with milk first and which with tea first. The crux of the example, for the question in point, is Fisher's insistence on a program of randomization for determining which of the eight cups get one treatment as opposed to the other.

Fisher's justification for the randomization falls into two categories. One sort of argument, which applies especially to this case but not generally, has to do with the fact that the object of the study is herself a rational being with the ability to react to information about the test, if made available to her. In this sense the test takes on the appearance of a game, since if 'told' which cups were treated with milk-first, the lady will react appropriately and vitiate the experimental aspect of the whole set-up. Therefore, in an attempt to keep information from her, the determination of which cups get the milk-first treatment is made through the randomizer. This precludes her guessing which cups were milk-first on the basis of knowing the investigator's biases or intentions. However, since most inquiries involve Nature, and not rational beings, this use of randomization to prevent the publicity of the experimental pattern is usually pointless. Therefore, let us consider the second sort of justification.

A more epistemologically serious matter involves the use of the randomizer to bypass the ignorance of the experimenter. Fisher argues like this: Suppose the lady's claim false, i.e. the null hypothesis true. It is desirable then to

replace the null hypothesis by a probabilistic one, say, that of the 70 possible divisions of the 8 cups (into two groups), of which only 1 (or 2, depending on the exact claim made by the lady) represents the presence of discriminatory powers, the frequency for each combination is constant, i.e. 1/70. Then the null hypothesis that the lady's claim is false is equivalent to a simple multinomial distribution with uniform frequency 1/70 to each possible division.

The random variable for the test is just the number of correctly identified cups (out of 8). But, we see that the multinomial probability distribution (given the denial of the lady's claim) is not forthcoming merely by putting four of the eight cups into one treatment group and the other four into the second. The first stage in obtaining a probability is to supply a *stochastic* basis. This is accomplished by use of the randomizer, which assigns an initial frequency of 1/70 to any division of the 8 cups into two groups. The second stage is to convert this frequency into a probability, even after seeing the outcome of the randomizer, i.e. even after knowing which cups get milk-first. Suppose, says Fisher, that in fact the lady's choice is determined completely by the set of factors fixed by four of the cups, e.g. the thick rimmed ones, or the the heaviest four, etc. So long as it is not known how these factors act on her choice, it does not matter which of the four cups are chosen by the randomizer for milk-first treatment. Of course there will be differences from one cup to another, even uncontrollable ones, but so long as these differences are not *known* to prejudice the lady's choice, knowledge of which cups were used for treatment will not interfere with the probability statement generated from the frequencies obtained through the randomization.

Kyburg reacts to Fisher's policy of randomization with a strong statement:

> I shall now argue that, in relation to perfectly possible and plausible states of our rational corpus, the simple precaution in question is neither a sufficient nor a necessary condition of the validity of the test of significance contemplated. It is not a sufficient condition. Suppose that four of the eight cups were thinner than the others, and suppose that we have reason to believe that a cup with a thinner edge is more likely to be judged as having been made with the milk first than a cup with a thicker edge. Now, if our randomization procedure is a suitable one, the four thin cups will receive the milk-first treatment one time out of seventy. But we perform the experiment only once; what do we do if our randomizing procedure does just happen to pick out these four cups for the milk-first treatment? Do we take another whirl with our randomizing apparatus? This will vitiate the sound frequency basis of our randomization, and will in general lead to the question of how biased our randomization has to be, and in what respects it is biased, before we reject a physically determined selection of cups. Or do we simply accept the rather odd result of our randomizing procedure, and admit that the experiment did not turn out well? In either case, it is clear that the mechanical randomization by itself is not sufficient to guarantee the reasonableness of the result of the experiment.[17]

Before considering the argument against necessity, note that Kyburg's charge against the sufficiency of randomization for avoiding bias is valid only given *the knowledge of how the factor in question,* e.g. thickness of the cups, *affects the lady's choice.* Under this condition, it is unreasonable to claim that randomization takes the place of a more carefully designed experiment, in which the thickness of cups is properly controlled. However, I argue that Fisher's use of randomization is to cover for *ignorance,* not *laziness.* In this regard, Kyburg's program supports Fisher's conclusion. If it is not known how thickness affects the lady's choice, then knowing that the randomizer selected the thick ones to get milk-first does not prevent the probability statement based on the frequencies from the randomization. That is, the strength rule intercedes to rule in favor of the randomness statement based on the initial knowledge, prior to knowing how the cups were divided. Restricting the reference set of divisions to those in which the thick cups get milk-first (and assuming the denial of the lady's claim) yields only degenerate [0, 1] frequencies for correct discrimination (since it is not known how the thickness operates on the lady's selection).

Continuing with Kyburg's criticism,

> The randomization of the treatments is not necessary, either. Let us suppose that the eight cups are really quite indistinguishable to us. We arrange them in a linear order, and apply the milk-first treatment to cups numbered 1, 2, 5, and 7. This is a perfectly possible result of the randomizing procedure; if the test of significance is valid when this selection of cups is produced by a randomizing procedure, it is valid when this selection is made arbitrarily. The experiment is only performed once.[18]

It is to be remembered that, for Kyburg, the frequencies used in the evidential basis (of the randomness statement) do not have any stochastic underpinning. Therefore, if the 1/70 frequency can be derived from a different reference class, generated by, say, repeated divisions at the hands of the investigator, then the randomization is not necessary. In this, Kyburg is correct. However, just where the 1/70 frequency for correct discrimination (given h_0) comes from in this case is mysterious. It cannot be merely the proportion of correct responses within the set of possible answers, since the set of possible answers does not automatically qualify as a reference set. Therefore, for Kyburg's argument to be valid, there must be some extra information, to wit: that the reference class generated by division at the whimsy of the investigator (and given h_0) leads to a 1/70 frequency for correctly picking the milk-first group. In the absence of this additional information, the randomization is necessary to supply the initial frequencies, i.e. to supply a reference class with precise frequencies.

It is to be emphasized that Kyburg's overall reaction to randomization is a very accurate one. The epistemological conditions under which it becomes a valid tool are severely restricted; no doubt more so than actual practice warrants. Therefore, his admonitions are in order. In the following chapter, dealing with Fisher's theory of relevance and its pertinence for experimental design, the keynote theme is found in this remark of Kyburg's, with which he opens his discussion of randomization:

> R.A. Fisher, in various of his publications, and particularly in *Statistical Methods for Research Workers*, and *The Design of Experiments*, shows most impressively the advantages to be gained from designing experiments in such a way as to take advantage of the knowledge we already possess – e.g., regarding the homogeneity or heterogeneity of different parts of the population we are investigating.[19]

The theme for the following discussion of Fisher's views on questions of experimental design takes this observation as the focal point, and tries to show that properly constructed experiments make full use of the background knowledge to restrict the interference of 'nuisance' factors. Where the background information leaves off, there Fisher opts for randomization to take up the slack. Thus, design trades off restrictions stemming from what is known against randomization over what is uncertain. Unfortunately, when randomization is given its proper due, especially with respect to relevance conditions, it is found to be incompatible with fiducial inference and our demands on principles of inductive relevance.

NOTES

[1] H.E. Kyburg, Jr., *Probability and the Logic of Rational Belief* (Middletown: Wesleyan University Press, 1961); H.E. Kyburg, Jr., *The Logical Foundations of Statistical Inference* (Boston and Dordrecht: D. Reidel, 1974).

[2] I. Levi, 'Direct Inference', *Journal of Philosophy*, LXXIV, #1 (1977), 5–29.

[3] Kyburg, *The Logical Foundations of Statistical Inference*, p. 207.

[4] Kyburg points this out in his article, 'Logical and Fiducial Probability', *Bull. Inst. Int. Stat.* 40[2] (1963), 884–901, see p. 894.

[5] Kyburg, *Logical Foundations of Statistical Inference*, pp. 227–28. These are merely heuristic principles which are refined in the randomness definition, pp. 233–235.

[6] This assumes that there is no peculiar knowledge about the unit class.

[7] *Ibid.*, p. 274.

[8] *Ibid.*, pp. 315–16.

[9] Kyburg, *Probability and the Logic of Rational Belief*, pp. 258–60, pp. 270–5, pp. 286–9.

[10] *Ibid.*, p. 288.

[11] Kyburg, *Logical Foundations of Statistical Inference*, pp. 165–9.

[12] Levi's counterexample is as follows:
Notation: $\%\,(B:A) = r$ for The Percentage of A's that are B's is r.
Example: Consider the evidential sentences:

e: $\%\,(B:A) = .9$. f_1: $\%\,(B:A \cap C) = .85$. f_2: $\%\,(B:A \cap C) = .92$.

f_3: $\%\,(B:A \cap C) = .95$. f: f_1 or f_2 or f_3.

g: x is a member of $A \cap C$. h: x is a member of B.

The following are legitimate Kyburgian confirmations, based on the foregoing:

$c(h/e\,\&\,g) = .9$ $c(h/e\,\&\,g\,\&\,f) = .9$

$c(h/e\,\&\,g\,\&\,(f_1 \text{ or } f_2)) = .9$ $c(h/e\,\&\,g\,\&\,(f_1 \text{ or } f_3)) = .9$

$c(h/e\,\&\,g\,\&\,f_1) = .85$ $c(h/e\,\&\,g\,\&\,f_2) = .92$

$c(h/e\,\&\,g\,\&\,f_3) = .95$

Levi shows that no coherent confirmation function satisfies the seven conditions (above). From correspondence between Levi and Kyburg, December, 1972. Also, appearing in Levi's 'Direct Inference', *Journal of Philosophy*, *op. cit.*, p. 21.

[13] See, for example, Fisher, 'Theory of Statistical Estimation', pp. 705–06.

[14] These bounds can be moved towards 0 and 1 by using a different non-sufficient statistic in place of m, but with higher correlation with $\bar{x}$, i.e. a smaller variance or greater efficiency than m.

[15] Kyburg, *Logical Foundations of Statistical Inference*, pp. 286–313.

[16] R.A. Fisher, *The Design of Experiments*, 8th ed., (New York: Hafner Publishing, 1971), pp. 11–26.

[17] Kyburg, *Probability and the Logic of Rational Belief*, p. 295.

[18] *Ibid.*, pp. 295–6.

[19] *Ibid.*, p. 294.

CHAPTER 8

RELEVANCE AND EXPERIMENTAL DESIGN

Thus far, the text establishes a common failure in three reconstructions of fiducial inference (as presented in Chapters 4, 6 and 7). In each case, the argument cannot defend a unique representation of 'ignorance'; however, an 'ignorance' assumption is a required premise of the argument. Nonetheless, we have uncovered an interesting difference between the positions taken by Hacking and Kyburg. The point of controversy can be located in the principles of relevance for direct inference. Hacking requires a proof of irrelevance by his Likelihood Principle. Kyburg's regulations governing randomness permit irrelevance unless relevance can be shown. Although neither works for fiducial theory, only Kyburg's formulation permits any justification of randomization (as a tool for experimental design). The goal in this chapter is to reconstruct the Fisherian views on experimental design by concentrating on conditions of relevance, 8.1. The upshot is that randomization cannot be salvaged, 8.2.

The treatment of relevance adhered to in the investigation of Fisherian inference (Chapters 3 to 5), is based upon the derived conditions of sufficiency and ancillarity (1.3), which stem from confirmational conditionalization (1.2). This approach proves essential for the proper handling of multi-parametric inference, especially in the step-by-step method (5.1). Historically, this defense of relevance conditions differs from Fisherian methods. In 8.1, I suggest an alternative path for developing relevance principles, based on Fisher's work in the theory of *estimation.*

The interpretation of estimation proposed in the text emphasizes the role of an estimator as a summary of a large body of data. That is, instead of seeing an estimate as a 'guestimate' (derived from given data) at the true, but unknown value of some quantity, the estimate is evaluated in terms of its adequacy as a replacement of the data for the purpose of inference about the unknown quantity. If we add to this an insight (in a form due to R. Hilpinen) that the Total Evidence requirement has two distinct (but related) interpretations, to wit: an obligation to use all the relevant data already acquired (TRE_1) and an obligation to seek out all the relevant evidence available through inquiry (TRE_2), then problems in design can be linked with problems in estimation through common standards of relevance.

For the purpose of the discussion in Chapters 3, 4 and 5, it sufficed to identify extremes in the range of possible estimators (now, viewed as summaries of data). A sufficient statistic summarizes all the relevant data present, whereas an ancillary statistic reports only irrelevant features of the data available. However, in reconstructing Fisher's theory of estimation, we find a more finely grained division in which estimators are ranked between the two extremes of *sufficient* and *ancillary.* Several criteria are suggested for constructing this ranking, e.g. efficiency and sampling variance properties of statistics, but in the end it is Fisher's concept of *Information* which is used to compare different estimators.

This Information measure fits nicely into the picture that design and inference are related through the Total Evidence principle(s). Fisher's Information measure can also

be applied to evaluate alternative designs. The text illustrates this application in *The Moldy Bread* example.

In the subsequent Section, 8.2, we conclude the discussion of randomization. The point to be made is that randomization has no merit on an 'after-trial' analysis. If the sampled entities are *epistemically homogeneous* after inspection, then randomization is irrelevant. If the sampled entities are *epistemically heterogeneous* after inspection, then having selected them by use of a randomizer is without benefit to an investigator whose worry is over bias in his sample. Moreover, in the latter case the problem may be avoided if, knowing the factors which lead to the heterogeneity, the investigator *balances* the samples. Thus, we see that overt randomization in experimental design is, at best, irrelevant to the inferences drawn once the data are available, and otherwise misleading, i.e. the pre-trial expectations based on randomization in sampling procedures can be significantly different from post-trial expectations.

The text ends with a summary, Section 8.3, reviewing the principal arguments and conclusions of the book.

8.1. FISHERIAN THEORIES OF RELEVANCE AND EXPERIMENTAL DESIGN

In the last two chapters we have examined Hacking's and Kyburg's theories and have emphasized their treatments of inductive relevance. We argued that they could not combine (1) a reconstruction of fiducial inference, with (2) a theory of general inverse inference, without altering the relevance conditions. Hacking's reconstruction of the fiducial argument, although only partially successful, rests on a strict principle of irrelevance which must be abandoned when the problem of 'typicalness' (p. 178) is confronted. Kyburg's system provides a handsome reconstruction for general inference from sample to population at the cost of the Sufficiency principle for fiducial inference. That is, his reconstruction of the fiducial argument is too free with the notion of 'rationally representative' samples (p. 192).

It is appropriate, therefore, to consider Fisher's analysis of inductive relevance with the hope that new insight into a theory of inverse inference might result. The position argued for in this chapter is that Fisher viewed the two problems (1 and 2, above) as being closely related and that he tried to solve both with the same machinery. Towards this end he developed a novel theory of relevance, attacking the issue from two perspectives: the significance approach and the likelihood approach. The two courses merged through the concept of Information, which led to the now standard concepts of Sufficiency and Ancillarity.

The reconstruction offered in Chapters 4 and 5 used these relevance principles and defended them from the Bayesian perspective. This was done to keep clear the ties to CC and coherence, especially for the difficult problem

of sufficiency in multivariate cases. In this section, using Fisher's ideas on Information, we reestablish these relevance conditions by relying on a more radically frequentist perspective. The *leitmotif* for the analysis is the search for a unified theory of relevance for inverse inference.

By way of introduction to the position attributed to Fisher, reconsider the TRE requirement. A non-rigorous formulation of the rule of total evidence requires that the investigator use all the relevant evidence available. Besides the serious question of criteria for inductive relevance, there is another ambiguity in the TRE principle. The ambiguity hangs on a double reading of 'availability'. It has been pointed out in the philosophic literature as recently as 1970 that there are at least two distinct senses for the TRE requirement.[1]

Under one reading, the one considered throughout the discussion of the fiducial argument, [TRE_1], the investigator is constrained to use all the evidence in hand. That is, he must extract all the relevant information contained in the *BK*. It is with respect to this reading that the restrictions on the pivotal variable are defended.

An alternative analysis of equal importance yields the interpretation, [TRE_2], that the investigator is required to *seek out*, through inquiry, the relevant evidence. When one remembers that scientific methods involve acquiring information through experimentation, and with particular emphasis on the correct design of the experiment, then the second reading of TRE carries great force. It says in part that, within the confines of practical limitations, e.g. cost in time and money, it is incumbent on the investigator to select an experimental procedure which will net a maximum of relevant evidence. Certainly the choice of a chance set-up, of a sampling rule to fix how the set-up is to be used, and also the method for reporting the outcomes (the statistic), all fall under the heading of experimental design.

It requires only a cursory reading of Fisher's work to recognize that, if anything, he was overly sensitive to the inherent duality of the TRE requirement. His first two books, *Statistical Methods for Research Workers* and *The Design of Experiments* has as a principal goal the resolution of general statistical problems through the integration of experimental concerns with parallel concerns facing the investigator, now qua statistician, who would be called in for the *post mortem:* that is, to help analyze the data. Fisher's awareness of this strategy, at least in retrospective, is adequately shown in his Presidential address to a 1961 International Symposium:

> When, a little more than 25 years ago, I first attempted a systematic exposition of the subject, known as the Design of Experiments, it is no very grave confession to avow that I did not fully understand the position among the statistical sciences of this new discipline.

My approach at that time was frankly a technological one. As a statistician I had often set myself the task of analyzing experimental data, and was much concerned with those improvements in statistical methods, which promised to make such analysis, more thorough and more comprehensive. Technically, I could see that some methods were superior to others in the concrete sense of extracting from the data more 'information' on the subject under enquiry, and therefore of leading to estimates of higher precision, and to tests of significance of greater sensitivity. And so it was, in this atmosphere, borne in upon me that very often, when the most elaborate statistical refinements possible could increase the precision by only a few per cent, yet a different design involving little, or no additional experimental labour, might increase the precision two-fold, or five-fold or even more, and could often supply information in addition on relevant supplementary questions on which the original design was completely uninformative.

It was thus clear at an early stage that there were quantitatively large technological gains to be obtained through the deliberate study of Experimental Design, and that these gains were to be harvested by making the plan of experimentation and observation logically coherent with the aims of the experiment, or, in other words with the kind of inference about the real world, which it was hoped that the experimental results would permit.[2]

The technical term 'Information' appears in numerous places in Fisher's writings. Often it plays a central role in the more theoretical ones. The position presented here is that, under conditions to be stated, Fisher's measure of 'Information' [Inf.] is also a *measure* of inductive relevance. Exactly what is meant by a 'measure' will be made clearer later on. Of major philosophic importance is the consequence that Inf. provides, in conjunction with the fiducial argument, a handle for dealing with both aspects of TRE.[3]

Consider first a heuristic argument which links TRE_2 to TRE_1, relating to the relevance conditions of significance testing. A commonplace fact of the laboratory is that the measurement of an object by a particular measuring apparatus constitutes a chance set-up which, to a good approximation, satisfies the normal distribution. For example, if one uses an ordinary thermometer to take a sequence of readings of an object at a fixed temperature, then the readings will (usually) be distributed in accord with the normal law. If the mean μ of the distribution equals the true temperature of the object, then the thermometer is said to be *unbiased.* Among unbiased thermometers, the ranking by different levels of accuracy, i.e. the precision of the instrument, is solely a matter of σ^2, the variance. The smaller the variance the higher the precision and the greater the accuracy of an arbitrary reading.

Thus, we note that the relative precision of the thermometers carries a comparative judgement about the quality of the data in particular measurement pairs. It seems intuitive that a measurement taken with a thermometer of greater precision supplies more information about the true temperature

than does a reading from a less precise thermometer. Continuing with the heuristic argument, given a choice between using two different unbiased thermometers, and discounting the cost for application, it again seems intuitive that the investigator should choose the more precise instrument. The performance for the more accurate measurement reflects a judgement in relevance under TRE_2.[4]

With the heuristic argument as a (primitive) guide to judgements about relevance, one can reconstruct a major part of Fisher's theory of statistical estimation, interpreted as a theory of relevance (under TRE_1), as it applies to significance testing. Suppose several readings, $x_1, \ldots, x_n$ (n odd), a sample, have been taken using an unbiased thermometer with a standard deviation, σ, of one unit. A fundamental problem facing the investigator is how to represent the sample data in a compact form, easily understood and capable of statistical evaluation. Specifically, if a significance test is to be run with null hypothesis h_0 positing a value for T^* the true, unknown temperature, then the reduction of the data must be one which conveys all (and only) relevant information about T^* contained in the original sample.

For estimation, Fisher defines a *consistent* statistic as

DEFINITION: A function of the observed frequencies which takes the exact parametric value when for these frequencies their expectations are substituted.[5]

If the instrument is unbiased, then by definition, a statistic of the data from the instrument is consistent for the true temperature T^*, just in case it is a consistent statistic for the mean μ of the hypothetical normal distribution.

Consider two alternative reductions of the data. One to the median m of the sample, and the other $\bar{x}$, the sample mean or sample average. It is a common feature of many statistics, i.e. reductions of the data, that in large samples they have the property of themselves satisfying a normal distribution. That is, $\bar{x}$ or m, calculated from a sequence of samples (large n) are distributed according to a normal law. More significantly, it can be shown that the means of these hypothetical distributions, i.e. $\mu_{\bar{x}}$ and μ_m, coincide with μ, the mean of the original distribution (which equal T^*, in this case, by assumption). Thus, both $\bar{x}$ and $\bar{m}$ are consistent statistics for hypotheses about T^*.[6]

However, the variances generated by the proposed statistics, i.e. $\sigma^2_{\bar{x}}$ and σ^2_m, are not equal. For instance, in a sample of size n, $\sigma^2_{\bar{x}} = \sigma^2/n$; whereas $\sigma^2_m \simeq (\pi/2) \cdot (\sigma^2/n)$. The use of the $\bar{x}$ statistic is thus comparable to a single reading from an unbiased instrument with variance $1/n$ ($\sigma^2 = 1$). The use of the m statistic is comparable to a single reading from an unbiased instrument with variance $(\pi/2) \cdot (1/n)$, which is greater than $1/n$. By the heuristic argument,

the instrument with greater precision (smaller variance) is preferred; it yields a greater amount of relevant evidence [TRE_2].

The corresponding criterion for TRE_1 is *efficiency* which measures the sampling variance of the competing consistent statistics. Hence, $\bar{x}$ is preferred to m as a statistic (for the significance test) since it carries more of the relevant evidence about T^* contained in the original sample.

Another intuition about information and relevance is that the data contain only restricted amounts of information. Their relevance is limited. It is well known that in some situations, e.g. those arising in the normal distribution, there exists statistics which, by the traditional (Bayesian) principles of relevance, convey *all* the relevant evidence about some parameter present in the original sample. These are, of course, just the *sufficient* statistics, and are used in place of the sample data without fear of losing relevant evidence through the substitution.

It has been shown that sufficient statistics have a theoretically minimum (limiting) variance among a large family of 'regular' estimators (asymptotically normal and unbiased with independent trials) for a wide range of distributions.[7] Moreover, with small samples, all efficient estimators (in Cramér's sense) are also sufficient estimators, although the converse is *not* true.[8] For several common distributions the two concepts, i.e. small sample efficiency and sufficiency, agree.[9] However, for Fisher the small sample question is handled by his Information concept [Inf.] (discussed below), which he develops from the large sample efficiency of sufficient statistics.[10]

In *Statistical Methods and Scientific Inference* (pp. 148–158) Fisher offers a somewhat technical argument in support of his Inf. as what I am calling a measure of inductive relevance. To paraphrase his analysis, let t be a consistent statistic of some data d for the parameter θ. Fisher emphasizes the need for the general consistency of t, regardless of the value of θ. Subject to this condition, the limiting variance of t can be minimized if t is based on the likelihood function $\mathrm{p}(d/\theta)$. $\mathrm{Inf}(t)$, defined by the recriprocal of the sampling variance of t, is maximized if t is based on the likelihood function $\mathrm{p}(d/\theta)$.

The final step in the argument is the construction of sufficient statistics from Informational considerations. The key to the solution is the observation that sufficient statistics permit the combination of independent samples without loss of relevant evidence. Therefore, it is required that the sum of $\mathrm{Inf}(t_1)$ and $\mathrm{Inf}(t_2)$ (t_1 from data d_1 and t_2 from data d_2) must equal the Informational content of the total evidence, (d_1, d_2), just in case t_1 is sufficient for d_1 and t_2 is sufficient for d_2.

This will occur whenever 'the likelihood function of the statistic $[\mathrm{p}(t/\theta)]$

shall be the same as that of all samples from which such a statistic could have been calculated.'[11] That is, using Information, the statistic t preserves all the relevant information whenever the likelihood function $p(t = k/\theta)$ is the same as the likelihood $p(d/\theta)$, for all data d satisfying $t(d) = k$. This is an equivalent definition of Sufficiency to the definition offered in Section 1.3. Hence, Fisher's Inf. measure assigns highest ranging to sufficient statistics, where they exist. By a similar argument, ancillary statistics are ranked lowest, i.e. if t' is ancillary then $\text{Inf}(t') = 0$.[12] Thus Information can be used as a relevance measure, under TRE_1, to identify sufficient or ancillary statistics, where they exist.

An interesting, open problem is the evaluation of the Information ranking for statistics that fall between the extremes of ancillarity and sufficiency. Is there any sense to be made of the ordering of statistics by their Information level? A few brief remarks will show that a simple answer is not available. First, except for the extremes, statistics with equal Information values need not be equivalent. That is, just because $\text{Inf}(t_1) = \text{Inf}(t_2)$ (t_1, t_2 based on the same data), does not mean that $p(t_1/t_2)$ is independent of the parameter of interest. Still more upsetting is the observation that even when Information levels are unequal for two statistics both may contain relevant information. In other words, even though $\text{Inf}(t_1) > \text{Inf}(t_2)$ it may still be the case that $p(t_2/t_1)$ depends upon the parameter of interest. Thus, a rule which dictates that the statistic with the lower Information rating be ignored for inference does not replace the TRE_1 principle. However, if one were faced with the unusual problem of using only one statistic as a substitute for the original data, such a rule would offer a plausible solution.[13]

The problem of a forced choice among statistics is somewhat artificial, if all are based on the same data. But a central theme in this discussion has been the search for a unified theory of relevance, to include both TRE_1 and TRE_2 concerns. Such a theory, it is argued, must evaluate the complete data-gathering/data-processing statistical procedure. The first phase, data-gathering, includes the selection of an experimental design. The second phase, data-processing, includes the summarization of the sample evidence. Information can be used to evaluate statistics, not merely as products of the second stage (a matter of TRE_1), but also as products of the choice at the first stage (a matter of TRE_2). It is a real (and not artificial) problem to have to choose a single experimental design. If the second phase is optimized through sufficient statistics, so that no better summaries of the experimental results are possible, then differences in the Information rankings of different statistics reflect Informational differences in the designs. Hence, Inf. can be used to measure

the worth of experimental designs in terms of the evidence they produce. In short, Information is used as a relevance measure for TRE_2. The remainder of this section is devoted to illustrations of this procedure.

The accompanying examples all make use of the Student's t-pivotal for deriving fiducial probabilities about unknown means of normal distributions when the variances also are unknown. In Section 5.1 we saw that the fiducial 'marginal' posterior probability for hypotheses about an unknown normal mean (in the absence of knowledge of the variance) is given by this pivotal (see p. 143). The missing factor in our analysis is the Information ranking for data $(\bar{x}, s^2)$ with respect to the parameter μ (the mean). Fisher gives this as:

$$\text{Inf}(\bar{x}, s^2) = [(v+1)/(v+3)] \cdot (n/s^2);^{14}$$

where s^2 is the sample variance and $v = n - 1$, the number of degrees of freedom for the variance estimate s^2, with n independent observations. As a reminder, $t = (|\bar{x} - \mu_{\bar{x}}|) \div (s/\sqrt{n})$.

For a first example, let us consider an overly simplified situation.[15] Suppose that in testing a new strain of wheat two independent trials, crops from two different fields, show a net increase of 8 and 9 bushels/acre. The average increase, $\bar{x}$, is 8.5 bushels/acre, and the corresponding estimated standard deviation for $\bar{x}$ $(s/\sqrt{2})$ is .5. Using $t = |8.5 - \mu_{\bar{x}}| \div .5$ $(v = 1)$, the fiducial probability that the new grain has an improved yield, i.e. $\mu_{\bar{x}} > 0$, is .98. If, instead, the two trial crops had shown a net increase of 8 and 18 bushels/acre, so that the sample average $\bar{x}'$ is increased to 13 bushels/acre, the new sample variance (for $\bar{x}'$) is altered to a value of 25, i.e. $s'/\sqrt{2} = 5$. Using the new data the fiducial probability for the hypothesis that the grain is an improved one is now only about .87.

The apparent anomaly, increased average yield yet decreased support for the hypothesis that the strain is improved, is explained away by correctly interpreting the variance. It is important to remember that the use of the normal distribution is justified only by empirical facts. These facts can be loosely summarized by two simple-minded causal laws:

(1) Affecting the growth of any particular seedling are numerous independent causal factors, e.g. soil fertility, the lighting intensity, atmospheric temperature and humidity, not to mention all those factors attributed to the genetic make-up of the grain.

(2) *Ceteris paribus*, similar causes yield similar effects.

The larger variance, therefore, represents the effect of an increased number of uncontrolled causal factors. There is an obvious analogy. The precision of an instrument can be compared to the degree of experimental control by

comparing the magnitude of variance of the statistics used to summarize this experiment. More generally, since Information gives a measure of precision, Information can be used to measure the degree of experimental control in a design: which is a measure of TRE_2.

When we apply the Information measure to the jointly sufficient statistics used in the t-pivotal we see that it is a function of two factors:

(*i*) the primary concern, the magnitude of the estimated variance, s^2, and

(*ii*) the number of observations used to construct this estimate.

As expected, the Information ratings for the two experiments are unequal, 2 and 1/50 respectively. Thus, we might well conclude that the data from the second field experiment carries comparatively little relevant evidence for hypotheses about the difference in quality of the grains and that in future tests the first field is preferred since it offers greater control.[16]

The Moldy Bread Example

Suppose a chemist goes to a large bread manufacturer, e.g. the Better Bread Co., with the claim that he has discovered a new treatment to prevent spoilage. The treatment allegedly increases the bread's resistance to contamination from air-borne bacteria. Also, the new method is cheap and does not alter either the taste or nutritional value of the bread. Conceivably, it might be agreed that if after testing the treatment is shown to be an effective spoilage retardant, Better Bread buys the rights for say $50,000.

A small scale version of the crucial experiment is: pick, arbitrarily, 8 loaves and separate them into two groups of 4. One is for treatment and the other is left untreated to act as a 'control'. After treatment, slice all the breads and select (randomly) one slice from each loaf (so there are 8 slices, 4 treated, and 4 untreated). Expose the slices to room air for a fixed time, 36 hours, and then compare the spoilage by weighing the bacterial colonies that develop. If the probability is greater than .99 that the treated bread has greater resistance, then Better Bread buys the process.

The experiment is run simultaneously by three laboratory workers: Walker, Jones and Smith. The common background assumptions are: The 4 untreated slices represent a sample of 4 from a normal population with unknown mean μ_u and unknown variance σ^2. Similarly, the 4 treated slices represent a sample of 4 from a normal population with unknown mean μ_t and the same unknown variance σ^2. The hypothesis that the treated bread is superior is equivalent to the hypothesis that $\mu_u - \mu_t > 0$. The t-distribution can be applied to a hypothetical parameter μ_{u-t}, the mean of the difference between the two types of bread. The corresponding variance σ^2_{u-t} is also unknown. Sufficient statistics exist for both parameters.[17]

Walker performs the experiment and calculates, from his data, that the fiducial probability for the hypothesis $\mu_{u-t} > 0$ is greater than .99 (See accompanying table.) Using Walker's data Better Bread buys the contract.

Suppose, now, that the second lab worker, Jones, reports to the management with his story. Jones says, "I used the same experimental design as Walker and obtained the same numerical data. (See accompanying table.) However, when I examined the slices of bread used in my test, I found that, by coincidence they came from parallel parts of the original loaves. That is, there was a middle slice among the treated sample slices and a middle slice among the control group slices, a left heel and a left heel, etc. Since all the loaves were sliced on the same machine, I thought I ought to use the extra bit of information."

Jones' story gives the statistician extra data. The obvious question is whether it is relevant. Under usual statistical procedures the extra information is used to pair the slices. That is, the claim that the two sets of data are independent, i.e. $\rho = 0$, is no longer secure. The original data are reformulated in a new table and the sequence of differences calculated. The sequence of differences can be used, again with the t-distribution, to calculate fiducial support for the hypothesis about μ_{u-t}, based on the set of statistics displayed in the table.[18] Using Jones' data, and the technique of 'paired differences', the fiducial probability that the treated bread is superior is approximately .985, which is less than the required .99. Since the probability for the hypothesis has changed by incorporating the extra information into the analysis, it is relevant. However, does the fact that Jones' experiment contains more (and different) information automatically make it a better experimental procedure? Does Jones' modified experiment represent an improved design?

Analyze the problem from the perspective of precision and control. Jones' report permits the introduction of control by the slicing machine. If the blades of the slicing machine are contaminated and retain their exact contamination throughout the 8 slicing operations, then it seems reasonable to conclude that the 8 slices of bread are not uniformly sterile at the start of the 36 hours. By comparing the parallel slices, the four pairs, Jones can eliminate the variation in initial contamination, introduced by the slicing machine and thus *increase* the overall precision. However, if the slicing machine is sterile, or worse yet, if its blades do not maintain a constant contamination between operations, then the pairing process only adds to the imprecision by emphasizing the differences between parallel slices (introduced, now, by the slicing machine and not the differing treatments).

TABLE I
(Walker's data)

untreated slices	treated slices
36.20 mg/cm²	30.25 mg/cm²
42.40 mg/cm²	27.50 mg/cm²
40.60 mg/cm²	29.75 mg/cm²
39.20 mg/cm²	27.80 mg/cm²

$\bar{x}_u = 39.600$ $\bar{x}_t = 28.825$
$s_u^2 = 6.85$ $s_t^2 = 1.90$
$\bar{x}_{u-t} = 10.775$ $s_{u-t}^2 = (s_u^2 + s_t^2) = 8.75$ $n = 4$ $v = 6$
(from the pooled variance)
$t_1 = |\bar{x}_{u-t} - \mu_{u-t}| \div (s_{u-t}/\sqrt{n}) = |10.775 - \mu_{u-t}| \div \sqrt{2.188}$

TABLE II
(Jones' data)

untreated	treated	difference
36.20	30.25	5.95
42.40	27.80	14.60
40.60	27.50	13.10
39.20	29.75	9.45

$\bar{x}_{u-t} = \bar{x}_d = 10.775$ $s_d^2 = 15.02$ $n = 4$ $v = 3$
$\mu_{u-t} = \mu_d$
$t_2 = |\bar{x}_d - \mu_d| \div (s_d/\sqrt{n}) = |10.775 - \mu_d| \div \sqrt{3.756}$

TABLE III
(Smith's data)

untreated	treated	difference
36.20	27.50	8.70
42.40	30.25	12.15
40.60	29.75	10.85
39.20	27.80	11.40

$\bar{x}_{d'} = 10.775$ $s_{d'}^2 = 1.65$ $n = 4$ $v = 3$
$t_3 = |\bar{x} - \mu_{d'}| \div (s_{d'}/\sqrt{n}) = |10.775 - \mu_{d'}| \div \sqrt{.4375}$

Inf. is sensitive to the appropriate factors. It rates Walker's data with a value of .355. Using Jones' modified procedure (with a reduced number of degrees of freedom) the Inf. value is only .178. Hence, using Inf. as a measure of relevance [TRE_2], the first procedure is preferred. Despite its less refined design, it maintains a greater level of precision. Note that part of the advantage of the first procedure (from design through inference) stems from the doubling of degrees of freedom. That this, in and of itself, is not the deciding factor, can be shown by considering the third laboratory worker, Smith, and his story.

Smith reports, "I too, ran the experiment using Walker's technique and obtained the same data. However, it happened that the 8 petri dishes chosen in my test came in pairs from 4 different sterilizers. By coincidence, the dishes were divided so that each of the two groups of bread slices used 4 differently sterilized dishes. It is possible that the sterilizers were not all of the same quality, and so I took the extra information into account."

When this is done (by matching, as with Jones' data), the fiducial argument yields a probability in excess of .99 for the hypothesis that the treated bread is superior. Information ranks Smith's procedure (from design through inference) with a value of 1.62, much greater than either of the others. Hence, his design is preferred over the two previous ones.

One important point must be emphasized. This use of Inf. is aimed at TRE_2 primarily, not TRE_1. Jones' *cannot* improve his analysis by disregarding the information about the slicing machine in the hope of duplicating Walker's original experiment and inference. Once the evidence is in hand TRE_1 takes over. However, it is with respect to design, before the data are in hand, that the use of Inf. to rate future activity is valid.

An interesting sidelight to this analysis is that it shows the limits of the advantages to be gained from randomization. After the data are in and the outcome of the randomizer is observed, TRE_1 takes over and, as will be discussed in 8.2, the extra information is (comparable to Jones' extra information) *relevant* and undesired. However, one might plan ahead of time to make sure that the outcome of the randomizer is forever kept a secret from the investigator. Jones' might plan for the future by making sure that he never gets to see where the slices come from. This is a potential solution, in favor of randomization, but with the obvious tasteless consequence that blinders are vindicated. In fact, this very strategy has been suggested by I.J. Good.[19]

To return to the Moldy Bread example, the key question is whether or not a particular design which plans for paired matching is desirable. What deter-

TABLE IV

number of paired observations	ρ for matched individuals (critical values)
2	.167
3	.160
4	.143
5	.127
6	.114
7	.103
8	.0935
9	.0856
10	.0790
11	.0733
12	.0683
13	.0640
14	.0602
15	.0569
26	.0351
50	.0188
250	.0049
500	.0020

mines the worth of the matching is the degree of added control it gives the experiment. This, in turn, is determined by two factors: the original population variance σ^2 and the correlation ρ between the matched individuals. If the number of pairs to be taken is known (4 for the bread example), then the question whether a particular pairing plan is worth adding to the design can be answered. For comparison with Walker's design ($\rho = 0$), which does not use matching, Table IV shows the critical (minimal) values for ρ (at a fixed sample size) at which a design with pairing becomes preferable. Note that the variance σ^2 is irrelevant to this decision.

Usually, the correlation for a pairing scheme is unknown. Therefore, the experimenter cannot tell ahead of time which set-up is preferable. Afterwards, the experimenter can make inferences about ρ (but not fiducially!). A simple decision rule, based on a *pilot* sample, is: Compare the Information value of the observed data with a hypothetical sample (differing only by the assumption that $\rho = 0$) and adopt the design with pairing just in case the Inf. value of the *pilot* sample is greater. This amounts to using Table IV, with the *pilot* sample correlation coefficient in place of ρ.

However, if ρ is known, then a set of jointly sufficient statistics exist and the problem becomes much simpler. As shown in Appendix 9.7, the full

$2(n-1)$ degrees of freedom can be achieved even with pairing. The resulting decision rule for selecting a design with matching is also simple. For any sample size and any variance, an experiment using a paired design is superior to one using independent samples if and only if the correlation (from the pairing) is positive.

The Moldy Bread example, despite its simplicity, reveals the importance of utilizing all the background information for evaluating alternative experimental designs. The difference between the two unknown means μ_t and μ_u can be tested by two different procedures. One employs a scheme for matching to restrict the samples. Control is added by using pairing to eliminate variations due to common properties. The alternative procedure uses independent samples, with an advantage due to increased degrees of freedom. The problem created by this trade off parallels the more complex problems arising for testing differences in (unknown) means of several kinds of treatments. The selection of designs available range from unrestricted 'fully randomized' samples, through partially controlled 'randomized block' patterns, to even more precise 'Latin Square' designs.[20] In the brief discussion which follows, I argue that the advantages offered by the increase in control depends upon the corresponding background assumptions, similar to the assumptions about the correlation from the pairing in the Moldy Bread example. Again, Information is a function of the relevant factors involved. Finally, the discussion ends with a short statement on the value of randomization in experimental design.

8.2. RANDOMIZATION AND RELEVANCE

In the previous chapter, where Kyburg's randomness conditions were discussed, I argued that randomization in design (in connection with Fisher's Lady Tasting Tea experiment) is supposed to provide the grounds for replacing uncertainty about the possible effects of nuisance factors with a probability statement about error. We noticed that, in Kyburg's theory, if the background information is rich enough to discriminate statistical *differences* among the randomized outcomes (according to Kyburg's randomness rules), then knowing which configuration has been selected by the randomizer interferes with the probability claim about experimental error; thus, the apparent advantage netted from the randomization is only illusory. In this section I argue that when the background information is detailed enough to defeat randomization in this fashion, then one can construct an improved design with added control and less randomization.

Let us begin with a summary statement of the problem. The background information includes knowledge of the effects of some factors present in the experimental set-up. There remain other nuisance factors, possibly controllable, whose causal relation to the objects of the test is unknown. It is not even known whether the nuisance factors are relevant. The observed variations from object to object are due to various combinations of factors of both sorts. Increased precision is obtained by controlling or eliminating factors. Using the knowledge about factors in the first group, one can restrict the experiment. Where background knowledge leaves off randomization over the nuisance factors takes up the epistemic slack. The more that is known about the experimental set-up the less the need for randomization. This can be made clearer by considering a simple example (basically, a linear model), which we can connect to our study of the Moldy Bread example.

Suppose two experimental treatments for property F are to be tested for comparison of their F-effects. Call them T_1 and T_2. Four subjects are to be randomly selected, say by number tables, with T_1 to be applied to two of them and T_2 to the remaining pair. The background assumptions include: F is a continuous magnitude, like weight. F is determined by numerous factors, some independent and some interrelated. In the population at large these factors combine so that F-magnitudes tend to be normally distributed, but the mean and variance of this hypothetical distribution are unknown. Introducing T_1 or T_2 adds one new independent factor for F, not already present in the population at large. It is believed that T_1 effects every subject identically, either increasing or decreasing the observed F-magnitude by the amount δ_{T_1} and similarly for T_2 with consequent change δ_{T_2}. Thus, if all subjects were administered T_1, the unknown population mean would be altered by the amount δ_{T_1} (possibly zero), and the variance would remain unchanged.

If μ represents the unknown mean, and σ^2 the unknown variance, then the difference between treatment$_1$ and treatment$_2$ can be determined from the difference between the two hypothetical populations: one given T_1 with mean $(\mu + \delta_{T_1})$ and variance σ^2, the other given T_2 with mean $(\mu + \delta_{T_2})$ and variance σ^2. Hence, the difference is $\delta_{T_1} - \delta_{T_2}$; which is just the comparison desired.[21]

Assume, first, that the background knowledge is weak so that very little is known about F, or the causal factors that influence its occurrence (magnitude). Under these conditions, the sample of four objects is *epistemologically* homogeneous for F, prior to measurement. In other words, additional information about the sample cannot serve to individuate (prior to measurement) differences in F-magnitudes. With the chance randomization this epistemological

indifference can be converted into a probabilistic indifference, by direct inference. The selection of two, by *chance*, for the T_1 group grounds the claim that they form an independent sample of two from a $N([\mu + \delta_{T_1}], \sigma^2)$ distribution. With epistemological homogeneity, this claim remains valid even if the outcome of the randomizer becomes known; since knowing the sample's numbers provides no *useful* information about F-magnitudes. Using the randomizer in this way and with the assumption of epistemological homogeneity, we find that the problem reduces to an instance of a bivariate normal distribution with $\rho = 0$.[22] As in the Moldy Bread example, the t-pivotal permits fiducial inference about the difference $\delta_{T_1} - \delta_{T_2}$.

Next, modify the background knowledge so that instead of near total ignorance about the causes of F, one knows that certain environmental factors, say diet, affect F-magnitudes. That this epistemological change may invalidate the previous analysis is evident when one imagines the following: The randomizer is used to select the samples, but the T_1 pair are both vegetarians while the T_2 pair enjoy at least one meat meal per day. Now, the claim that the samples represent the $N(\mu + \delta_{T_i}, \sigma^2)$ distributions $(i = 1, 2)$ is no longer valid, even though the samples are selected randomly, i.e. by the randomizer. The information that the diets are radically different is relevant. The populations of individuals with restricted diets, either vegetarians or meat eaters, have, we suppose, different normal distributions for F-magnitudes.

From the assumptions about the causes of F, since a restriction on diet (eating meat or abstaining from it) limits variations in some relevant factors causing F, the variance for the resulting populations are decreased. This change increases the significance of any observed difference noted between the two treatments, just as if precision is increased. That is, the t-pivotal is a ratio whose denominator is a function of the variance. The problem is that the restriction by diet may also affect the unknown population mean. Then the difference between the sample means reflect not only the difference in treatments $\delta_{T_1} - \delta_{T_2}$, but also the difference in diets! If, say, meat eaters tend to be F-ier than vegetarians, then this difference is added to whatever advantage T_1 provides, undermining the assessment that observed differences are a result of causal differences in the treatments only.

The increase in background information which shows diet relevant to F-magnitudes, besides invalidating the first design, permits an improved design which takes diet into account. The sampling pattern can be restricted so that there is a balance between the two groups and precision is added. For example, the samples might be stratified to include matching diets across paired members. This design improvement corresponds to the benefits of

using Smith's experiment in place of Walker's experiment (in the Moldy Bread example).

The paired designs we have considered form a special instance of 'block' designs, which are based on a one-way classification among different treatment groups (one of which may be labelled the 'control'). By balancing the design we can eliminate heterogeneity while increasing precision. In every case there is a trade off between the addition of control by matching against the loss of degrees of freedom for the estimates of variance needed. As in the Moldy Bread example, the exchange can be measured by Fisher's Information function. Unless the matching fixes a significant causal factor (with the linear model, unless the correlation is high for matched pairs), the restricted design is disadvantageous compared to an unrestricted design.

It is to be noticed that randomization is still used in 'block' designs, but in a limited way. With matching, the division of the samples is determined so that each treatment group includes subjects of each kind. In the example above, one vegetarian is included in each group. Which vegetarian goes to which group is determined *randomly*. Relative to the modified background information the vegetarians form an epistemically homogeneous group, just as the meat eaters do. Of course an even further enriched background might invalidate the randomized block design. For instance, it might be known that genetic make-up is relevant to F-magnitude. The 'block' design does not prohibit close relatives from being selected for a common treatment group, once diet is fixed.

By a strategy similar to that used to incorporate information about diet, a two-way restriction ('Latin Square') can be constructed to reflect the knowledge that genetics are also relevant. Again there is another trade off of added control for loss of degrees of freedom in estimates of variance. For example, with two treatments, each added control cuts the degrees of freedom in half.[23] A quick look at Table IV (p. 213) shows that the restrictions cannot be piled up heedlessly, since with small degrees of freedom the magnitude of the correlation is a serious factor. Thus, with a moderately rich background it is a non-trivial question which (relevant) factors to restrict. The problem is even more serious in tests among several treatments.

The sequence of restrictions from fully randomized samples to highly stratified designs reflects the changes in the background knowledge which accompanies each stage. Paralleling the added restrictions is the elimination of randomization. Historically, the debates over the advantages of each design often revolved about the randomization issue. A good case in point is the argument between Fisher and Gossett ('Student') whether there is any benefit in

randomizing.[24] Although I doubt that either of these distinguished men would disagree with the conclusion that the merits of a design depend upon the background information available to the investigator (discounting cost factors), their debate missed this point and became hung up on less important matters.

For instance, Fisher's standard response to anyone who argued for totally balanced designs (without any randomization), was to demonstrate that in repeated trials on the same 'field' the estimates would invariably be biased (one way or the other) if the non-randomized plan were reused. Not so for the randomized designs.[25]

In response, it was pointed out (by Gossett, for one) that on any particular trial the randomized plan might lead to the same configuration as in the fully balanced design.[26] A 'fully random' design might also be 'random block' design which might also be a 'Latin Square' design, on a particular trial. This criticism is related to the stopping rule problem (p. 49), since the design determines the sample space of possible configurations and a given pattern may be a member of any number of different such spaces. I believe that Gossett's retort points out the inadequacy of Fisher's response.

However, using the results of our previous investigation, we can offer a substitute for Fisher's answer. Instead of worrying about the bias of an estimate on repeated trials with a given design, it could be argued that balanced designs are inappropriate whenever the background knowledge is insufficient to warrant the added control. That is, until the background information includes the relevance of the factors fixed by the balanced design, there is no justification for the selection of a more restricted scheme. In fact, as with some matching plans in the Moldy Bread example, unnecessary limitations may violate TRE_2!

Discussions of Fisher's contributions to the theory of experimental design often concentrate on his insistance on randomization. We have tried to show that correct assessment of the value of different designs (as measured by Information) depends not on randomization, but on the background knowledge available prior to the experiment. The final problem, then, is to determine whether Fisher's program for using randomization on nuisance factors is successful.

It is not, for two reasons. Let us resume the investigation of relevance principles which apply to the complete data-gathering/data-processing procedure. Fisher's objections to randomized tests for inference (p. 153) can be redirected at randomization in design. Randomizers, no matter whether they are used to determine rejection regions or configurations for experimental designs, yield outcomes which are ancillary for the hypotheses tested. Hence,

by TRE_1 relevance considerations, randomizers cannot improve a design, once the correct limitations (balance, matching, etc.) are set.

In the examples presented in this section randomization served to ground direct inference. If epistemological homogeneity obtains, then (it was argued) the statements of direct inference retain their validity even after the outcome of the randomizer becomes known. Is this a reasonable conclusion? Suppose an investigator is ignorant of the effects of factor G on the properties tested. Imagine that subjects bearing G are randomly assigned to the various test groups. What if all those with G land in one test group? If the investigator is puzzled before the randomization about the bias resulting from the G factor, is he any less puzzled after the randomization? In order to subdue this puzzlement, so that it does not interfere with direct inference arguments before randomization, it is necessary that questionable cases of relevance be analyzed as Kyburg analyzes them: *relevance must be demonstrated*! Then, the outcomes of the randomization do not upset direct inference, since (by epistemic homogeneity) the nuisance factors cannot be shown to be relevant.[27]

The second difficulty, then, is that relevance has to be analyzed one way for randomization and a different way for fiducial inference. Regarding fiducial inference, we argued that a strict irrelevance principle is needed (irrelevance must be demonstrated). Therefore, our unpleasant conclusion is that Fisher's program for inverse inference does not allow a unified theory of relevance.

8.3. SUMMARY

What have we learned from our investigations? Which problems have been solved and which remain open for future inquiry?

To begin with, our picture of fiducial theory is much clearer. We can safely say that, with respect to the conditions of adequacy formulated in Chapter 1, fiducial inference is invalid. The program falters over the very issue which motivates its construction. That is, frequentists like Fisher object to the Bayesian solution(s) to inverse inference on the grounds that the prior probability adopted for the solution is unsupported by a direct inference and has no clear ties to empirical (stochastic) probability. Put succinctly, some alternative to traditional Bayesian inference is needed when the agent is relatively ignorant about the unknown parameters which are the subject of his inquiry. The five adequacy conditions (of Chapter 1) entail that whatever inference is substituted for the Bayesian one, it must admit a Bayesian model,

i.e. it must cohere with the probability calculus and obey confirmational conditionalization. Frequentists are prepared to sanction direct inference, so the obvious try is to reduce problems of inverse inference to solvable versions of direct inference. Formally, the locus for the shift is the pivotal variable. Prior to obtaining data, direct inference about the pivotal is straightforward. By assuming the data are irrelevant to this inference one carries out the reduction, since given the data probability about the pivotal is logically equivalent to probability about the unknown parameter.

All the steps in the argument can be made rigorous, except there is no solution which avoids using a privileged representation for the unknown factor(s). Moreover, which representation is selected for the honor is contingent upon irrelevant features of the data (irrelevant according to the standards set by the five conditions of adequacy in 1.2). In other words, the fiducial argument has a Bayesian model whose prior probability component acts as the 'ignorance' prior. But this prior probability is *not* unique with irrelevant changes in the data. Hence, fiducial theory fails to provide an acceptable representation for the very sort of ignorance which drives Frequentists to object to alternative Bayesian solutions.

We have seen how Fisherian inference is hierarchical: significance testing is the base since it applies in situations of greatest ignorance and its conclusions are informatively weakest. Likelihood based inference sits mid-way between significance testing and fiducial theory. The latter, though applicable in rare circumstances only, supplies conclusions with the logical force of probabilities; hence, it is the most sophisticated inference pattern of the three. My reconstruction (Chapters 3 to 5) ties fiducial theory to significance testing (and envelopes likelihood through the D_p ranking – see 3.2). Therefore, the faults found in fiducial theory (4.8) filter down the hierarchy and have consequences for significance testing and likelihood inference too. These problems have not been examined in detail in this book (except for the version of likelihood theory depicted by Hacking's program and discussed in Chapter 6). Certainly, there is room for additional research along these lines to clarify the limitations in significance testing and likelihood theory which arise from the failures in fiducial inference.

In the text we have considered two alternative programs, due to Ian Hacking and Henry Kyburg (Chapters 6 and 7), and I argue that neither offers a satisfactory resolution to the problems for fiducial theory. However, the two differ in ways that lead us to serious questions about inverse inference viewed outside the fiducial argument. By focusing on the differences in the relevance conditions for direct inference, we see an interesting issue open up in the

theory of experimental design; namely, whether randomization can be made consistent with established principles of relevance, e.g. sufficiency and ancillarity.

In order to answer this question, our discussion turned to an examination of Fisher's radical views on estimation and design (Chapter 8). These two topics were joined through the dual readings of the total evidence principle: one must use all the evidence already collected [TRE_1] and also one is obliged to seek out the most valuable information available [TRE_2]. By giving estimation theory an uncommon interpretation (an estimate *summarizes data*), we uncover a novel use for *Information*: as a lynch pin holding TRE_1 and TRE_2 together. Thus, criteria for judging the adequacy of an estimate as a summary of data, i.e. criteria for judging relevance under TRE_1, also serve as criteria for assessing alternative designs as sources of relevant evidence, i.e. criteria for assessing relevance under TRE_2. The upshot is that Fisher's arguments for randomization, all of which are constructed on a 'before-trial' basis, do not stand up to the force of our 'after-trial' demands. These 'after-trial' demands are, of course, nothing more than the TRE_1 relevance conditions forthcoming from the theory of estimation.

Our analysis deals Fisherian inference two distinct blows. [A] Fiducial theory is incoherent. [B] The Fisherian doctrine which insists on overt randomization in experimental design is unsound. Where does this leave us en route to a satisfactory theory of statistical inference? The first result reminds us that, with respect to inverse inference, ignorance is not bliss – for we have no means of taking advantage of our ignorance through a fiducial argument while respecting the canons of the inductive logic described in Chapter 1. But the second result brings good news! If, as I have tried to argue, randomization is not crucial to proper inductive methodology, then inverse inference need not be restricted to situations where the data are collected under the 'protection' of a randomizer. We do *not* have to bolster inverse inference by designing randomized studies. This opens the field to observational studies, e.g. retrospective studies, where randomization is impossible. It is exciting to realize that it is a mistake to classify case-histories as second rate investigations solely because, say, randomization is not the method of division used to separate treatment and control groups. The outstanding challenge, then, is to devise legitimate procedures for inverse inference, procedures which do not rely only on the 'before-trial' direct inferences involving the randomizers (which inferences do not survive the 'after-trial' analysis).

One beginning point for such a project is with the problem of finding

adequate summaries of data from observational studies. Despite the faults in the Fisherian program, the approach to problems of relevance through estimation and Information remains sound and productive. It is most fitting that, perhaps, what we can learn about inductive relevance from R.A. Fisher will shed light on our methods of obtaining and evaluating statistical data relating to matters of public health.

NOTES

[1] R. Hilpinen, 'On the Information Provided by Observations', in Hintikka and Suppes eds., *Information and Inference*, (Boston and Dordrecht: D. Reidel, 1970), pp. 97–122.

[2] R.A. Fisher, 'The Place of the Design of Experiments in the Logic of Scientific Inference', *Sankhyā* 27–28 (1965–66), 33–8.

[3] In this and in the following discussion an important characteristic of the 'experiments' studied is the large degree of control presupposed. A consequence of this assumption is the ability, on the part of the investigator, to choose among several experimental designs. For Fisher, those alternative designs which merit serious consideration all include some measure of randomization. In the next section I try to show that the importance of randomization is inversely proportional to the systematization incorporated; which, in turn, should be proportional to the richness of the background knowledge available. Moreover, I argue that to allow randomization to serve the role intended by Fisher, a marked change must occur in attitudes held towards demonstrating irrelevance (from the position suggested in the reconstruction of the fiducial argument, as presented in Chapters 4 and 5).

It would be desirable, therefore, to be able to carry out the proposals made here for unifying inference with design, without appeal to randomization. Unfortunately, the theory of experimental design is not very clear where the initial presumption of control is denied. Without significant amounts of control the opportunity for randomization is diminished. Hence, the subsequent investigation focuses on the more traditional cases, including randomization, only because the theory of design is already developed for such contexts. An important goal of this program, to be pursued in a future inquiry, is the evaluation of design and inference in observational studies where control is lacking. For a current review of the central issues involved see: S.M. McKinlay, 'The Design and Analysis of the Observational Study – A Review', *J. Am. Stat. Assoc.* 70, #351 (1975), 503–520.

[4] Inf., which will be analyzed in the approach from likelihood, captures these intuitions by selecting, in this case, $1/\sigma^2$ as a criterion for relevance. The reader is reminded that the heuristic argument is informal and serves only to motivate the following discussion. A good example of the breakdown in the connection between the sampling variance of a statistic and its Inf. rating is to be found on p. 460 of L. Savage's article, 'On Rereading R.A. Fisher', *The Annals of Statistics* 4, #3 (1976), 441–500.

[5] Fisher, *Statistical Methods and Scientific Inference*, p. 150.

[6] Fisher's first definition of consistency was not in terms of expected values, but in terms of limiting values of the statistic as the sample size became infinitely large. Specifically: "That the probability that the error of estimation exceeds in absolute value any

value ϵ, or, symbolically,

$$P(|T-\theta| > \epsilon) \text{ [where T is the statistic]}$$

shall tend to zero as the size of the sample is increased, for all positive values ϵ, however small." This definition was subsequently rejected in favor of that in the text because of the following interesting argument:

With respect to a function of the observations, T, defined for all possible sizes of sample this definition has a certain meaning. However, any particular method of treating a finite sample of N_1 observations may be represented as belonging to a great variety of such general functions. In particular, if T′ stands for any function whatsoever of N_1 observations, and T_n for any function fulfilling the asymptotic condition of consistency, then

$$(1/N)(N_1 T' + [N-N_1]\, T_{N-N_1})$$

is itself a statistic defined for all values of N, and tending asymptotically to the limit θ, yet it is recognizable when $N = N_1$ as the arbitrary function T′ calculated from the finite sample.

In fact, the asymptotic definition is satisfied by any statistic whatsoever applied to a finite sample, and is useless for the development of a theory of small samples; *ibid.*, pp. 150–151.

The philosophical importance of this argument applies to an interpretation of probability, not discussed. The justification H. Reichenbach gives to his 'straight rule' follows, exactly, Fisher's original argument for defending the limiting definition of consistency; Hans Reichenbach, 'The Justification of Induction', in Foster and Martin eds., *Probability, Confirmation and Simplicity* (New York: Odyssey Press, 1966), pp. 423–434. If we accept Fisher's own objection to the asymptotic definition, then Reichenbach's justification carries no weight for small samples.

[7] This follows from the asymptotic sufficiency of the maximum likelihood estimator and special asymptotic properties of the variance of the maximum likelihood estimate; Fisher, 'Theory of Statistical Estimation', in *CMS*, paper #11.

[8] H. Cramér, *Mathematical Methods of Statistics* (1946, Princeton: Princeton University Press, 1971), pp. 477–490.

[9] L.J. Savage, *The Foundations of Statistics*, 2nd ed., (New York: Dover, 1972), pp. 220–245. For a counterexample, see the reference in footnote 4 to Savage's recent article.

[10] Fisher, *op. cit.*, p. 709.

The notion of second order efficiency is due to C.R. Rao; see, for example, 'Criteria of Estimation in Large Samples', *Sankhyā* 25 (1963), 189–206. Recent discoveries due to J.K. Ghosh and K. Subramanyam establish the second order efficiency of the maximum likelihood estimator under each of two different measures of second order efficiency given by Rao; see 'Second Order Efficiency of Maximum Likelihood Estimators', *Sankhyā* 36 (1974), 325–58.

Intuitively, first order efficiency measures the *average* loss in Fisher Information that results when the data are replaced by some estimate. Second order efficiency measures the *total* loss in Fisher Information for the same replacement. In all cases the measurements are carried out for their limiting values as the sample size increases.

Rao bases his analysis on Fisher's two (early) papers: 'On the Mathematical

Foundations of Theoretical Statistics', *Philos. Trans. Roy. Soc., Ser. A* **222** (1922), 309–65; and 'Theory of Statistical Estimation', *Proc. Camb. Phil. Soc.* **22** (1925), 700–25. I have shown that, in addition to these sources, one can extract a measure of second order efficiency from a decomposition of χ^2 for contingency tables. This result is in agreement with Fisher's remarks in Section 57 of *Statistical Methods for Research Workers*.

[11] Fisher, *Statistical Methods and Scientific Inference*, p. 157.

[12] Most of this material can be found in several places in Fisher's writings. Besides the references cited in notes 10 and 11, see 'Two New Properties of Mathematical Likelihood'; *The Design of Experiments*, Chapter XI; *Statistical Methods for Research Workers*, Chapter IX; and most especially *Statistical methods and Scientific Inference*, Chapter VI.

[13] At least the statistic with the higher Information level must contain some relevant information not present in the lower ranked statistic. Call a statistic t' sufficient with respect to the family τ of statistics $(t_1, \ldots, t_n)$, for the parameter θ, just in case $\mathrm{p}((t_1, \ldots, t_n)/t')$ is independent of θ. Then $\mathrm{Inf}(t') > \mathrm{Inf}(t)$ for all t in τ which are not sufficient with respect to τ. For example, if τ is the pair (t_1, t_2) and t_1, but not t_2, is sufficient with respect to τ, then $\mathrm{Inf}(t_1) > \mathrm{Inf}(t_2)$.

[14] Fisher, *The Design of Experiments*, p. 244.

[15] Fisher, *Statistical Methods for Research Workers*, 14th ed. (New York: Hafner, 1973), pp. 126–7.

[16] A problem which, so far as I know, does not admit a fiducial solution is calculation of the posterior probability for the hypotheses based on *both* experiments. The complicating factor is the knowledge of a common (theoretical) mean for the difference between the new and old grains. A solution was offered by Fisher in an article; 'Sampling the Reference Set', *Sankhyā, Ser. A 23* (1961), 3–8. It had the exceptional feature of generating "simultaneous probability statements about parameters . . . not limited to cases of simultaneous exhaustive estimation"; *ibid.*, p. 8. This is in direct conflict with the 'step-by-step' procedure. However, it is reported by a student of Fisher, that "after publication Fisher regarded his expression as wrong (Cornish, personal communication)"; A.M. Verhagen, *The Notion of Induced Probability in Statistical Inference*, Technical Paper #21 (Melbourne: CSIRO, 1966), p. 20. Therefore, it will be supposed that no valid posterior fiducial probability for hypotheses about the common mean is available.

[17] This follows from the results (p. 151) for the bivariate normal. *BK* is such that the means are unknown, the variances unknown, but known equal, and the correlation is known to be 0.

[18] The legitimacy of this as a fiducial argument is questionable. If the new correlation were known, then a modified test would be in order (discussed below). However, with the correlation unknown, but the variances assumed equal, the step-by-step procedure is not available.

[19] I.J. Good, 'Random Thoughts About Randomness', in K. Schaffner and R. Cohen, eds., *PSA 1972* (Dordrecht: D. Reidel, 1974), pp. 122–4.

[20] Fisher, *The Design of Experiments*, pp. 50–74.

[21] Technically, this is a bivariate normal distribution with unknown means, unknown variances but variance ratio known to be one, and correlation yet to be determined. The parameter under investigation is the *difference* in the two means.

[22] The role of the randomization in this argument is to provide the stochastic (probability$_2$) element necessary for direct inference about the sample. If, by comparison, the

sample is selected as a result of the investigator's intuitions about which are representative samples, then no stochastic process is identifiable for the direct inference. This problem, as it arises for Kyburg's theory, is discussed on p. 198.

[23] Traditionally, data from these kinds of experiments are subjected to analysis of variance tests. With such methods the extra degrees of freedom are lost to the estimates of the variance of controlling factors.

[24] 'Student' (W.S. Gossett), 'Comparison Between Balanced and Random Arrangements of Field Plots', *Biometrika* 29 (1937), 363–79.

[25] See, for example, Fisher, *The Design of Experiments*, pp. 74–80.

[26] Gossett, *op. cit.*, p. 366.

[27] This differs from the situation in the Moldy Bread example in an important way. The matching or pairing was unforeseen in the Moldy Bread example. Neither Smith nor Jones anticipated that their *random* samples would reflect additional properties suitable for constructing a 'block' design. Once the data were obtained, however, since they were *unprepared* to dismiss as irrelevant the extra information, they could not continue with their original belief that the samples were epistemically homogeneous. They were forced to recast the data, by matching, so that a new claim of homogeneity might be sustained. This meant reworking the analysis just as if a 'block' design had been decided upon from the start. In the problems considered in this section, the question that arises is whether randomization can successfully eliminate the threat of bias when the sources of the bias (nuisance factors) are acknowledged at the start.

APPENDICES

9.1. PIVOTAL PROBABILITIES

The basis of the fiducial argument is direct inference about the pivotal variable. The pivotal variable is a hypothetical random variable defined in terms of the random variable to be observed, x, and the unknown parameter, θ. The special property of the pivotal variable on which the argument is based is that it has a known distribution. A simple proof of this is as follows.

Let $v = \mathbf{V}(x, \theta)$ be a pivotal variable. By definition, v has the same distribution $\mathbf{F}(v)$ for each value of θ in the set of values Δ, where Δ is provided by the model. From the background assumptions, Δ exhausts the set of possible values of θ. Hence $\theta^{\dagger}$, the true value of θ, is in Δ. Therefore, $\mathbf{F}(v)$ is the actual distribution of v.

It is worthwhile to consider two alternative arguments for this conclusion, neither of which is satisfactory. First, since for each θ in Δ $\mathbf{F}(v/\theta) = \mathbf{F}(v)$, one might try a Bayesian argument to derive an unconditional probability for v, $\mathbf{p}(v)$, by introducing a prior probability for θ, $\mathbf{p}(\theta)$. Then, by the multiplication theorem, $\mathbf{p}(v) = \int_{\Delta} \mathbf{p}(v/\theta) \cdot \mathbf{p}(\theta) \cdot d\theta$; where $\mathbf{p}(v/\theta)$ is obtained by direct inference from $\mathbf{F}(v/\theta)$. Since v is a pivotal variable, $\mathbf{p}(v/\theta)$ is independent of θ. Hence,

$$\mathbf{p}(v) = \mathbf{p}(v/\theta) \cdot \int_{\Delta} \mathbf{p}(\theta) \cdot d\theta. \qquad \text{[I]}$$

Now if the prior probability $\mathbf{p}(\theta)$ is *proper*, then the integral in [I] equals 1, so $\mathbf{p}(v) = \mathbf{p}(v/\theta)$ as desired. But, as the analysis in the text reveals, for most fiducial arguments the Bayesian model which duplicates the inference uses *improper* conjugate priors. In such cases the integral in [I] is unbounded and the argument, just presented, is invalid.

A second attempt which fails relies on a powerful inductive principle, Conglomerability,[1] due to de Finetti. Conglomerability is defined as follows (see p. 163):

Suppose that an event H can be partitioned into pairwise exclusive and mutually exhaustive events H_{α}, with α ranging over all cardinals less than some cardinal β. Let the conditional probability (density) $\mathbf{p}(E/H_{\alpha})$ satisfy the

inequality

$$k_1 \leqslant \mathrm{p}(E/H_\alpha) \leqslant k_2, \text{ for all } \alpha.$$

Then conglomerability asserts that

$$k_1 \leqslant \mathrm{p}(E/H) \leqslant k_2.$$

If v is substituted for E, θ partitions Δ which is substituted for H, then $\mathrm{p}(v) = \mathrm{p}(v/\theta)$. More generally, with Bayesian models using proper priors Conglomerability follows from confirmational principles already adopted; in fact the previous argument provides the proof. However, multivariate fiducial inference is incompatible with Conglomerability (see p. 164). Therefore, this argument too is unsatisfactory for deriving pivotal probabilities.

9.2. THE EQUIVALENCE OF FISHER'S AND TUKEY'S PIVOTAL CONDITIONS

It is claimed in the text (p. 111) that Fisher introduced his monotonicity/unboundedness requirement in order to limit fiducial arguments to 'natural' pivotals, i.e. pivotals which support coherent fiducial inference. To fascilitate the reconstruction, Tukey's 'smooth invertibility' condition has been substituted for Fisher's requirement. A pivotal is *smoothly invertible* (p. 111) if:

(*A*) It has the same range for any possible observation.

(*B*) It is 1-1 in both the random variable x and parameter θ.

(*C*) For continuous variables, this 1-1 mapping is continuous with continuous derivatives.

Fisher considered only continuous random variables when constructing pivotals (see p. 107) and he naturally assumed condition (*C*). It is reasonable to claim that Fisher's notion of unboundedness coincides with condition (*A*) of smooth invertibility. At least, that is how he seems to have used the constraint (see note 8, p. 134). It remains to show that, given (*A*) and (*C*), monotonicity and (*B*) are equivalent.

Let $v = \mathrm{V}(x, \theta)$ be a pivotal meeting conditions (*A*) and (*C*). Suppose that V is not monotonic in x for $\theta = \theta_1$. (I assume that monotonicity is to be understood as *strict* monotonicity.) Since V is continuous in x (from (*C*)), there must exist values $x_1 < x_2 < x_3$ in the neighborhood of x_2 (the region where monotonicity fails) such that $\mathrm{V}(x_1, \theta_1) = \mathrm{V}(x_3, \theta_1)$. But then V is not 1-1 in x for $\theta = \theta_1$. Similarly, if $\mathrm{V}(x_4, \theta_2) = \mathrm{V}(x_5, \theta_2)$ $(x_4 \neq x_5)$, then V is not strictly monotonic in x for $\theta = \theta_2$.

Tukey's 'smooth invertibility' condition offers the advantage of being applicable to discrete distributions simply by dropping (*C*).

9.3. COHERENCE OF ONE PARAMETER FIDUCIAL ARGUMENTS

The following is a sketch of a proof that fiducial arguments meeting the three conditions (below) are coherent. It is assumed that only one parameter distributions with sufficient statistics for all sample sizes are considered. This duplicates the assumptions used by D.V. Lindley in his article, 'Fiducial Distributions and Bayes' Theorem' (see note 18, p. 135).

Fisher states that the fiducial probability (using an unbounded/monotonic pivotal) is given as:

$$\mathbf{p}(\theta/x) = -\partial \mathbf{F}/\partial\theta;^{2}$$

where $\mathbf{F}(x, \theta)$ is the cumulative distribution function for x, given θ. The reconstruction in the text supports this fiducial probability just in case:

(i) There exists a pivotal variable defined in terms of x and θ, with distribution $\mathbf{F}'(v)$.

(ii) The pivotal is *smoothly invertible* (see Appendix 9.2).

(iii) The pivotal is *canonical* (see p. 125).

To show coherence it suffices to demonstrate that if such a pivotal exists, then the distribution can be transformed to location parameter form (see p. 24) since Lindley shows this to be both necessary and sufficient for coherence.

Let $v = \mathbf{V}(x, \theta)$ be a pivotal which satisfies (ii) and (iii). Then

$$\partial \mathbf{F}/\partial x = d\mathbf{F}/dv \cdot \partial \mathbf{V}/\partial x; \qquad [1]$$

where $|d\mathbf{F}/dv| = d\mathbf{F}'/dv$, i.e. the density function for v. Examine [1] at the special value $\theta = \theta^*$ for which v is canonical. Since $\mathbf{V}$ is monotonic (see Appendix 9.2), at $\theta = \theta^*$ $\mathbf{V}$ is decreasing for either decreasing or increasing values of x. If the former, then $v = x$. This is because with $\theta = \theta^*$ v and x have the same distribution, hence the same range, and (by smooth invertibility) $\mathbf{V}$ is 1-1 between x and v for constant θ. If the latter, then $\partial \mathbf{F}/\partial x$ is a symmetric function. Transforming v to v' by $\mathbf{F}'(v') = 1 - \mathbf{F}'(v)$ gives us another canonical pivotal v' which satisfies the first case. Hence, either $v = x$, in which case $\partial \mathbf{F}/\partial x = d\mathbf{F}/dv$, or else $x = v'$, in which case $\partial \mathbf{F}/\partial x = -d\mathbf{F}/dv$. Therefore,

$$|\partial \mathbf{V}/\partial x| = 1, \text{ evaluated at } \theta = \theta^*. \qquad [2]$$

Solutions to [2] are of the form

$$v = \mathbf{H}(\mathbf{R}(x) - \mathbf{S}(\theta)); \qquad [3]$$

for some functions **H**, **R** and **S**. Since v is a pivotal, its distribution is independent of θ. Therefore, transforming x to $\mathbf{R}(x)$ and θ to $\mathbf{S}(\theta)$ puts the problem in location parameter form.

This shows that the reconstructed fiducial argument is coherent for one parameter distributions involving data of two kinds (Section 4.4) or data of one kind which admits a common sufficient statistic (Section 4.5).

Conjecture: Where Fraser's *structural inference* provides a reconstruction of fiducial inference (see note 26, p. 26), then θ^* is the identity element of the structural group.

9.4. EXTENDED COHERENCE OF FIDUCIAL PROBABILITY

The purpose of this appendix is to show that both Hacking's and my own reconstructions are not limited to 'the variable to be observed.' It is reasonable that the fiducial argument should not be affected by transformations of the data which are 1-1, differentiable and *epistemically calculable* (see p. 89). Such transformations are information preserving. For example, transforming the data from x to x^3 should not alter the fiducial conclusions available. (That the transformation be 1-1 and differentiable is, of course, directly related to clauses (*B*) and (*C*) of smooth-invertibility – see Appendix 9.2.)

Let $\mathbf{T}: x \to x'$ be any transformation that satisfies the three conditions above.

Case$_1$: Hacking's reconstruction. If we assume that x supports a fiducial inference to $\mathbf{p}(\theta/x) = |\partial \mathbf{F}/\partial\theta|$, then Hacking has shown that

$$\partial \mathbf{F}/\partial x = \mathbf{G}(\mathbf{F}) \cdot \mathbf{H}(x), \qquad [1]$$

for some functions **G** and **H**.[3] However,

$$\partial \mathbf{F}/\partial x' = \partial \mathbf{F}/\partial x \cdot d\mathbf{T}^{-1}/dx'. \qquad [2]$$

Since $\mathbf{T}^{-1}$ is a calculable function of x', let $\mathbf{J}(x') = d\mathbf{T}^{-1}/dx'$. Clearly,

$$\mathbf{H}(x) = \mathbf{H}(\mathbf{T}^{-1}(x')). \qquad [3]$$

Therefore,

$$\partial \mathbf{F}/\partial x' = \mathbf{G}(\mathbf{F}) \cdot \mathbf{H}(\mathbf{T}^{-1}(x')) \cdot \mathbf{J}(x'). \qquad [4]$$

[4] can be rewritten as [1]. By appealing to Lindley's results, Hacking argues that [1] is also a sufficient condition for coherent fiducial inference.[4] Thus, replacing x by x' does not alter the fiducial conclusion about θ.

Case$_2$: Reconstruction using smoothly invertible, canonical pivotals. Let $v = V(x, \theta)$ be such a pivotal variable. From the results presented in Appendix 9.3,

$$\partial F/\partial x = |dF/dv|, \text{evaluated at } \theta = \theta^*; \qquad [5]$$

where θ^* is the special value for which v is canonical. Define v' as $T(v) = v'$. Therefore,

$$\begin{aligned} dF/dv' &= dF/dv \cdot dT^{-1}/dv' \\ &= dF/dv \cdot J(v'); \end{aligned} \qquad [6]$$

where J is defined in Case$_1$ (above). Since $v = x$ for $\theta = \theta^*$ (see Appendix 9.3), $v' = x'$ for $\theta = \theta^*$. Hence,

$$J(v') = J(x'), \text{for } \theta = \theta^*. \qquad [7]$$

Substituting in [2] using [5] and [7] yields,

$$\partial F/\partial x' = |dF/dv| \cdot J(v'), \text{evaluated at } \theta = \theta^*.$$

Hence, v' is canonical for x' with $\theta = \theta^*$ (from [6]). Coherence follows by the same argument used in Appendix 9.3, now with x' and v' replacing x and v.

9.5. HACKING'S IRRELEVANCE PRINCIPLE AND TUKEY'S SMOOTH INVERTIBILITY REQUIREMENT

Hacking's Likelihood Principle of Irrelevance [LPI] (see p. 173) is the irrelevance principle which permits retaining pivotal probability statements *after* observing the random variable. LPI is a composite rule incorporating three conditions: existence of a pivotal variable, existence of a partition induced by the pivotal variable (p. 173) and equivalent likelihood functions for pivotal support before and after observing the random variable. It is appropriate to compare these conditions with the restrictions used in the reconstruction supplied in Chapter 4: existence (1) of a smoothly invertible (2), canonical (3) pivotal. The restrictions are jointly equivalent for one parameter distributions with sufficient statistics, since they both amount to using fiducial inference just in case the problem can be converted to location parameter form.

In comparison, clause by clause, (*A*) of Tukey's smooth invertibility condition (p. 111) replaces Hacking's assumption that the pivotal induces the *same* partition before and after seeing the random variable. That the partition

has pairwise exclusive elements (in terms of the random variable and parameter) guarantees that the pivotal is a function of the random variable and parameter. Condition (*C*) of smooth invertibility (for continuous distributions) is tacitly assumed in Hacking's requirement that likelihood support for the pivotal is determined by *experimental* densities (pp. 175–177).

Next, condition (*B*) of smooth invertibility follows from LPI, as the following argument demonstrates. Assume, for simplicity, that the distributions $\mathbf{p}(x/\theta)$ are discrete. If $v = \mathbf{V}(x, \theta)$ is not 1-1, say $\mathbf{V}(x_1, \theta_k) = \mathbf{V}(x_2, \theta_k) = v_k$ $(x_1 \neq x_2)$, then since v is a pivotal variable, $\mathbf{p}(v_k) = \mathbf{p}(v_k/\theta_k)$. Therefore $\mathbf{p}(v_k) \geqslant \mathbf{p}(x_1/\theta_k) + \mathbf{p}(x_2/\theta_k)$. (Equality holds just in case x_1 and x_2 are the only solutions to the equation $\mathbf{V}(x, \theta_k) = v_k$.) By the LPI, x is irrelevant to v; hence, $\mathbf{p}(v_k/x_1) = \mathbf{p}(v_k)$. Also by the Law of Likelihood (p. 171) the likelihood for v_k given x_1, is (if defined) $\mathbf{p}(x_1/\theta_k)$. Therefore, by LPI $\mathbf{p}(x_2/\theta_k) = 0$, since substituting $\mathbf{p}(x_1/\theta_k)$ for $\mathbf{p}(v_k)$ in the inequality above, $\mathbf{p}(x_1/\theta_k) \geqslant \mathbf{p}(x_1/\theta_k) + \mathbf{p}(x_2/\theta_k)$. But by LPI, $\mathbf{p}(v_k/x_2) = \mathbf{p}(v_k)$. Therefore, by a similar argument, $\mathbf{p}(x_1/\theta_k) = 0$. Finally, $\mathbf{p}(v_k) = 0$, i.e. V is 1-1 in x over the sample space for v.

To show that V must be 1-1 in θ for fixed x, one need only remember that Hacking requires that, given x, $v = v_j$ must be equivalent to some statement $\theta = \theta_j$ or else his LPI does not apply.

9.6. HACKING'S SECOND FREQUENCY PRINCIPLE

Hacking's Frequency Principle$_1$ (p. 172) says that with evidence that the chances (probability$_2$ – **P**) for outcomes of trials of kind *K* are given by the distribution *D*, then the probability (probability$_1$ – **p**) for the hypotheses about outcomes of trials of kind *K* are given by distribution *D*. The Frequency Principle$_1$ is a rule for direct inference.

In order to have direct inference about pivotal kind of trials, another rule (stronger than the Frequency Principle$_1$) is needed. Hacking says,

> More generally suppose trials of kind K' are derivative on trials of kind K, and that whenever the distribution of chances on trials of kind K is in the class Δ, then the chance of E on a trial of kind K' is some definite value $\mathbf{P}_\Delta(E)$, greater than 0. Then
>
> $$\mathbf{p}(\langle X, K, \Delta; T, K', E\rangle / \langle X, K, \Delta; T, K', \Omega\rangle) = \mathbf{P}_\Delta(E).^5$$

This is what we have called the Frequency Principle$_2$ (p. 172). Hacking does not discuss this principle in any detail. It should be obvious that this rule supplies direct inference about the pivotal variable. That the chance $\mathbf{P}_\Delta(E)$ is well defined for the pivotal variable is arguable. The suggested proof in Appendix 9.1 is available for this very point. Furthermore, if countable

additivity is assumed, then the Frequency Principle$_2$ can be derived from the multiplication theorem, CC and a *proper* prior, just as in the arguments discussed in Appendix 9.1.

9.7. NOTES FOR THE BIVARIATE NORMAL DISTRIBUTION, INCLUDING A DERIVATION OF THE MASTER SIGNIFICANCE TEST [MST]

Notation: The bivariate normal $N(\mu, \Sigma)$ is such that random variables $\begin{pmatrix} x \\ y \end{pmatrix}$ have means $\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \mu$, and variance-covariance matrix

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} = \sigma_{yx} \\ \sigma_{yx} = \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \Sigma;$$

where

$$\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y.$$

In accord with Anderson, *An Introduction to Multivariate Statistical Analysis*,[6] define a linear transformation T of the random variables $\begin{pmatrix} x \\ y \end{pmatrix} \xrightarrow{T} \begin{pmatrix} z \\ w \end{pmatrix}$ by:

$$z = a_{11}x + a_{12}y \text{ and } w = a_{21}x + a_{22}y.$$

By *Theorem 2.4.4*,[7] $\begin{pmatrix} z \\ w \end{pmatrix}$ is $N(T\mu, T\Sigma T')$; where T is the transformation matrix:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

It is assumed that T is non-singular, i.e. $|T| \neq 0$. Also, T′ is the matrix whose elements a'_{ij} are the elements a_{ji} of T. T′ is called the *transform* of T.

LEMMA$_1$: Let T be any degenerate, non-singular transformation of $N(\mu, \Sigma)$, i.e. $a_{12} = a_{21} = 0$, $T \neq 0$, then $\rho_{xy} = \rho_{zw}$.

Proof:

$$T\Sigma T' = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \Sigma \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} =$$

$$\begin{pmatrix} a_{11}^2\sigma_x^2 & a_{11}a_{22}\sigma_{xy} \\ a_{11}a_{22}\sigma_{xy} & a_{22}^2\sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_z^2 & \sigma_{zw} \\ \sigma_{zw} & \sigma_w^2 \end{pmatrix}.$$

But $\rho_{zw} = \sigma_{zw}/\sigma_z\sigma_w = a_{11}a_{22}\sigma_{xy}/a_{11}a_{22}\sigma_x\sigma_y = \sigma_{xy}/\sigma_x\sigma_y = \rho_{xy}$.

COROLLARY: The Master Significance Test.

$$a(\mu_x - x) - b(\mu_y - y) \text{ is } N(0, a^2\sigma_x^2 + b^2\sigma_y^2 - 2ab\sigma_x\sigma_y\rho_{xy}).$$

Proof: Let $\begin{pmatrix} x \\ y \end{pmatrix}$ be $N(\mu, \Sigma)$.

Lemma. For constants c_1, c_2 the transformed variables $\begin{pmatrix} x - c_1 \\ y - c_2 \end{pmatrix}$ are distributed $N(\mu' \Sigma)$; where

$$\mu' = \begin{pmatrix} \mu_x - c_1 \\ \mu_y - c_2 \end{pmatrix}.$$

Proof: omitted.

Lemma. If $\begin{pmatrix} x \\ y \end{pmatrix}$ is $N(\mu, \Sigma)$ then $(x \pm y)$ is

$$N(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2 \pm 2\sigma_x\sigma_y\rho_{xy}).$$

Proof: Appealing to *Theorem 2.4.5*,[8]

Case_1 (+): Let $T+$ be the transformation defined by the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$.

Case_2 (−): Let $T-$ be the transformation defined by the matrix $\begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}$.

Note that μ_x and μ_y, although unknown, remain constants. Therefore, by the two lemmas and Lemma_1,

$$a(\mu_x - x) \pm b(\mu_y - y) \text{ is } N(0, a^2\sigma_x^2 + b^2\sigma_y^2 \pm 2ab\sigma_x\sigma_y\rho_{xy}).$$

Application: The Studentized version of the MST

Assume that the correlation and variance ratio are known. The purpose of this application is to develop estimates which have a full $2(n-1)$ degrees of freedom, based on n independent paired samples from a bivariate normal

distribution.[9] This corresponds to the second stage of the step-by-step argument for the bivariate normal distribution (p. 151).

The data consist of n independent, paired readings $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ $(1 \leqslant i \leqslant n)$, from $N(\mu, \Sigma)$. Let $c\sigma_x = \sigma_y$, for known constant c, and assume ρ_{xy} is known. To Studentize the MST it is required that a sufficient statistic exist for the generalized variance $(a^2\sigma_x^2 + b^2\sigma_y^2 \pm 2ab\sigma_x\sigma_y\rho_{xy})$. This corresponds to the identification of a sufficient statistic s^2 for σ^2 in the normal distribution (with μ unknown).

With c known, obviously a sufficient statistic for σ_x is also a sufficient statistic for σ_y. That is, there is only one free parameter in the covariance matrix Σ. It is common knowledge that if $\rho_{xy} = 0$ and $c = 1$, then a sufficient statistic for σ^2 is obtained by 'pooling' the estimates s_x^2 and s_y^2 (which are independent since $\rho_{xy} = 0$). (See, for example, Table I on page 211). With the use of a sequence of transformations, below, this technique is applicable for any correlation (other than $\rho_{xy} = \pm 1$) and variance ratio.

Let $\mathbf{T}_1$ be the transformation defined by the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1/c \end{pmatrix}; \text{ where } \sigma_y = c\sigma_x.$$

By Lemma_1: $\begin{pmatrix} x \\ y \end{pmatrix} \mathbf{T}_1 \begin{pmatrix} z_1 \\ w_1 \end{pmatrix}$ is $N(\mu_1, \Sigma_1)$; where $\mu_1 = \begin{pmatrix} \mu_x \\ \mu_y/c \end{pmatrix}$ and

$$\Sigma_1 = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho_{xy}/c \\ \sigma_x\sigma_y\rho_{xy}/c & \sigma_y^2/c^2 \end{pmatrix}.$$

Note that $\sigma_x = \sigma_{z_1} = \sigma_{w_1}$ and that $\rho_{xy} = \rho_{w_1 z_1}$.

LEMMA_2: Let the random variables x and y be $N(\mu, \Sigma)$ with $\sigma_y = c\sigma_x$. Let $\mathbf{T}_2$ be the transformation defined by the matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. Then the application of the transformation $\mathbf{T}_2$ to the variables x, y results in the variables z_2, w_2 which are $N(\mu_2, \Sigma_2)$. $\rho_{z_2 w_2} = 0$ just in case $c = 1$, i.e. $\sigma_x = \sigma_y$.

Proof: By *Theorem 2.4.4*. $\Sigma_2 = \mathbf{T}_2 \Sigma \mathbf{T}_2'$ which is the matrix

$$\begin{pmatrix} \sigma_x^2 + c^2\sigma_x^2 - 2\sigma_{xy} & \sigma_x^2 - c^2\sigma_x^2 \\ \sigma_x^2 - c^2\sigma_x^2 & \sigma_x^2 + c^2\sigma_x^2 + 2\sigma_{xy} \end{pmatrix}$$

But $\rho_{z_2 w_2} = \sigma_{z_2 w_2}/\sigma_{z_2}\sigma_{w_2}$. Substituting in Σ_2, above, it follows that $\rho_{z_2 w_2} = 0$ just in case $c = 1$. Also, it is interesting to note that this result is independent of the value of ρ_{xy}.

The application of $\mathbf{T}_1$ to the variables x and y results in the variables z_1 and w_1, which have the property that $\sigma_{z_1} = \sigma_{w_1}$. We may therefore use Lemma$_2$, and apply transformation $\mathbf{T}_2$ to obtain the independent variables z_2 and w_2. We then have the result that $\mathbf{T}_2 \circ \mathbf{T}_1 \begin{pmatrix} x \\ y \end{pmatrix}$ is $\mathbf{N}(\mu_2, \Sigma_2)$.

$$\mu_2 = \begin{pmatrix} \mu_x - \mu_y/c \\ \mu_x + \mu_y/c \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 2\sigma_x^2(1-\rho_{xy}) & 0 \\ 0 & 2\sigma_x^2(1+\rho_{xy}) \end{pmatrix}.$$

A final transformation, $\mathbf{T}_3$, eliminates the undesired factors $(1 \pm \rho_{xy})$, resulting in random variables z_3, w_3 which are independently distributed with common variance σ_x^2.

Define $\mathbf{T}_3$ by the matrix

$$\begin{pmatrix} (2(1-\rho_{xy}))^{-.5} & 0 \\ 0 & (2(1+\rho_{xy}))^{-.5} \end{pmatrix}$$

Hence, $\mathbf{T}_3 \circ \mathbf{T}_2 \circ \mathbf{T}_1 \begin{pmatrix} x \\ y \end{pmatrix}$ is $\mathbf{N}(\mu_3, \Sigma_3)$.

$$\mu_3 = \begin{pmatrix} (\mu_x - \mu_y/c)\cdot(2[1-\rho_{xy}])^{-.5} \\ (\mu_x + \mu_y/c)\cdot(2[1+\rho_{xy}])^{-.5} \end{pmatrix} \text{ and } \Sigma_3 = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_x^2 \end{pmatrix}.$$

Now a sufficient statistic for σ_x^2 is obtained from z_3 and w_3 by 'pooling' the variances $s_{z_3}^2$ and $s_{w_3}^2$. Let $s^2 = (s_{z_3}^2 + s_{w_3}^2)/2$. Then s^2 is the desired statistic with $2(n-1)$ degrees of freedom. Skipping some tedious calculations, we find that $s^2 = (s_x^2 + s_y^2/c^2)\cdot(1/2[1-\rho_{xy}^2]) - rs_x s_y \rho_{xy}/(c[1-\rho_{xy}^2])$; where s_x^2 and s_y^2 are the sample variances for the observed variables x and y, and r is the sample correlation coefficient.

Finally, the variance (denominator) of the MST can be rewritten as: $\sigma_x^2(a^2 + b^2c^2 + 2abc\rho_{xy})$. If s^2 is substituted for σ_x^2 in this formula and the results used as a denominator in a modified MST, then this modified MST has a t-distribution with $2(n-1)$ degrees of freedom.

A special instance which often arises for 'paired' testing is the case $c = 1$. Using the formula for s^2 (above), we find that the sufficient statistic for σ_x^2 is

$$s^2 = (s_x^2 + s_y^2)/2[1 - \rho_{xy}^2] - rs_x s_y \rho_{xy}/(1 - \rho_{xy}^2),$$

which is the familiar 'pooled' estimate:

$$(s_x^2 + s_y^2)/2,$$

for $\rho_{xy} = 0$. Also, the estimate (MST denominator) for the variance $\sigma_{(x-y)}^2$ (corresponding to the test for the difference $(\mu_x - \mu_y)$) is:

$$s_{(x-y)}^2 = (s_x^2 + s_y^2)/(1 + \rho_{xy}) - 2rs_x s_y \rho_{xy}/(1 + \rho_{xy}),$$

which is the familiar variance sum:

$$(s_x^2 + s_y^2),$$

for $\rho_{xy} = 0$.

9.8. ADDITIONAL NOTES FOR FIELLER'S (ALSO, THE MST) INTERVAL ESTIMATE OF THE RATIO OF MEANS IN A BIVARIATE NORMAL DISTRIBUTION

The purpose of this appendix is to show why Fieller's solution (also the MST solution), which is a confidence interval solution, is not a valid fiducial solution to the ratio of means problem.[10] Specifically, the smooth invertibility condition is violated by the MST pivotal. Both clause (*A*) and clause (*B*) of smooth invertibility fail with the MST pivotal.[11]

For simplicity, we display the failure for one case:

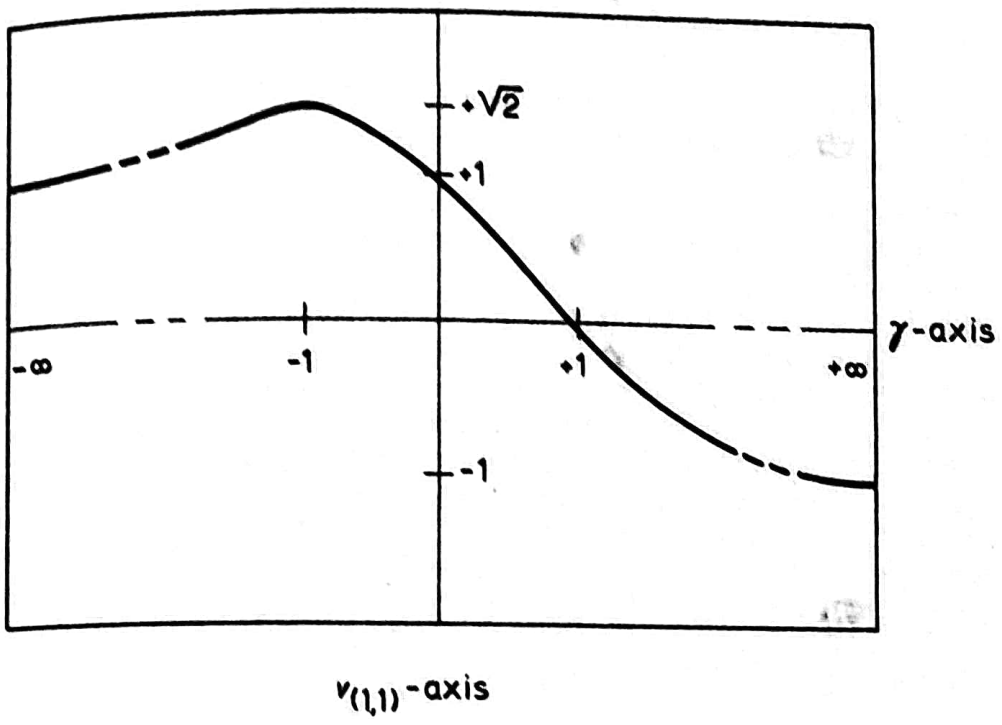
Let $\sigma_x^2 = \sigma_y^2 = 1$, $\rho_{xy} = 0$. Using the MST (p. 154), one can construct a pivotal variable for the unknown ratio $\gamma = \mu_x/\mu_y$.

Define v as $v = (x - y\gamma)/\sqrt{(1 + \gamma^2)}$.

v has a known distribution N(0, 1) for all possible ratios $-\infty \leqslant \gamma \leqslant +\infty$. It is interesting to note that this pivotal results from the MST by using an *incalculable* transformation (p. 89). The transformation relies on γ which is unknown.

Suppose the experimental data are $x = y = 1$. Substituting in v yields $v_{(1,\,1)} = (1 - \gamma)/\sqrt{(1 + \gamma^2)}$. However, $v_{(1,\,1)}$ is neither unbounded nor 1-1 between values of v and γ. Both these deficiencies are displayed in the graph of $v_{(1,\,1)}$ (below). It can be seen that $v_{(1,\,1)}$ is restricted to the interval: $-\infty < v_{(1,\,1)} \leqslant \sqrt{2}$, for all finite γ. (There are asymptotes at $v_{(1,\,1)} \to -1$ as $\gamma \to +\infty$, and at $v_{(1,\,1)} \to +1$ as $\gamma \to -\infty$. The maximum $(\sqrt{2})$ is reached at the value $\gamma = -1$.) Hence, v fails clause (*A*) of smooth invertibility.

Second, for any value γ', $-1 < \gamma' \leqslant 0$, there exists another value γ'', $-\infty \leqslant \gamma'' < -1$, such that



Graph of $v_{(1,\,1)}$: showing asymptotes; maximum point $(-1, \sqrt{2})$; and 1-1 failure for $\gamma < 0$.[13]

$$(1-\gamma')/\sqrt{(1+\gamma'^2)} = (1-\gamma'')/\sqrt{(1+\gamma''^2)}.$$

Thus, v is not 1-1 given the observations $x = y = 1$.

Confidence intervals are obtained by rejecting all γ such that the resulting v values are greater than k or less $-k$, i.e. $|v| > k$. However, $v_{(1,\,1)}$ is bounded. Therefore with the data (1, 1), all confidence intervals with $k > \sqrt{2}$ are logically trivial; encompassing all possible $\gamma \epsilon (-\infty, +\infty)$. This limiting k value corresponds to a probability level of approximately .85. Therefore, using $v_{(1,\,1)}$, all intervals of (at least) the .85 level are logically trivial. Also, for $1 < k < \sqrt{2}$ the rejection region consists of a set of ratios in the neighborhood of $\gamma = -1$, i.e. reject all γ such that $-1-\delta \leqslant \gamma \leqslant -1+\epsilon$ (δ and ϵ both small and positive). This results in a disconnected confidence interval. In other words, for k in this range (corresponding to a probability level between .85 and .67, approximately) the confidence interval is 'two-tailed': ratios are accepted iff they fall below $-1-\delta$, or if they fall above $-1+\epsilon$.

NOTES

[1] B. de Finetti, *Probability, Induction and Statistics* (London: John Wiley and Sons, 1972), pp. 98–107. de Finetti rejects Conglomerability because he does not accept

countable additivity, which it entails. Our program admits π-conglomerability (p. 167) but rejects Conglomerability because of its incompatibility with multivariate fiducial inference.

[2] Fisher, *Statistical Methods and Scientific Inference*, p. 74.

[3] Hacking, *Logic of Statistical Inference*, p. 155.

[4] *Ibid.*

[5] Hacking, *op. cit.*, p. 135.

[6] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis* (New York: John Wiley and Sons, 1958), pp. 11–3.

[7] *Ibid.*, p. 25. THEOREM 2.4.4: If $\mathbf{X}$ is distributed according to $N(\mu, \Sigma)$, then $\mathbf{Z} = \mathbf{DX}$ is distributed according to $N(\mathbf{D}\mu, \mathbf{D}\Sigma\mathbf{D}')$.

[8] *Ibid.*, p. 26. Theorem 2.4.5 generalizes 2.4.4. to include singular transformations D.

[9] It has come to my attention that these results were (probably) anticipated by E.A. Cornish, *Simultaneous Fiducial Distribution of Parameters*, Technical Paper #6 (Australia: CSIRO, 1960), as reported in A.M. Verhagen, *op. cit.*, pp. 20–1.

[10] E.C. Fieller, 'Some Problems in Interval Estimation', Symposium on Interval Estimation, *J. Royal. Stat. Soc., Ser. B* 16 (1954), 175–85.

[11] For additional discussion of Fieller's solution see J.S. Williams, 'The Role of Probability in Fiducial Inference', *Sankhyā, Ser. A* 28 (1966), 271–96, pp. 292–3.

[12] This graph duplicates Fieller, *op. cit.*, p. 178.

BIBLIOGRAPHY

Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc., 1958.

Anscombe, F.J., 'Dependence of the Fiducial Argument on the Sampling Rule', *Biometrika* 44 (1957), 464–9.

Anscombe, F.J., 'Tests of Goodness of Fit', *J. Royal Stat. Soc., Ser. B* 25–26 (1963–64), 81–94.

Barnard, G.A., 'Significance Tests for 2 × 2 Tables', *Biometrika* 34 (1946), 123–38.

Barnard, G.A., 'The Meaning of a Significance Level', *Biometrika* 34 (1947), 169–82.

Barnard, G.A., 'Some Logical Aspects of the Fiducial Argument', *J. Royal Stat. Soc., Ser. B* 25–26 (1963), 111–4.

Barnard, G.A., 'The Logic of Least Squares', *J. Royal Stat. Soc., Ser. B* 25–26 (1963), 124–7.

Barnard, G.A., 'Review of: *The Logic of Statistical Inference*', *Brit. J. Phil. Sci.* 23 (1972), 123–32.

Bartlett, M.S., 'Statistical Information and Properties of Sufficiency', *Proc. Roy. Soc., Ser. A* 154 (1936), 124–37.

Bayes, T., 'An Essay Towards Solving a Problem in the Doctrine of Chances', *Phil. Trans. Royal Soc.* 53 (1763), 370–418.

Bennett, G.W. and Cornish, E.A., 'A Comparison of the Simultaneous Fiducial Distributions Derived from the Multivariate Normal Distribution', *Bull. Int. Stat. Inst.* 40^2 (1963), 902–18.

Birnbaum, A., 'Concepts of Statistical Evidence', in S. Morgenbesser, P. Suppes, and M. White eds., *Philosophy, Science, and Method*. New York: St. Martin's Press, 1969.

Brillinger, D.R., 'Examples Bearing on the Definition of Fiducial Probability with a Bibliography, *Annals Math. Stat.* 33 (1962), 1349–55.

Buehler, R.J. and Feddersen, A.P., 'Note on a Conditional Property of Student's *t*', *Annals of Math. Stat.* 34 (1963), 1098–1100.

Carnap, Rudolf, *Logical Foundations of Probability*. Chicago: Univ. of Chicago Press, 1962.

Cox, D.R., 'Some Problems Connected with Statistical Inference', *Annals of Mathematical Statistics* 29 (1958), 357–63.

Cramér, Harald, *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 1971.

Creasy, M.A., 'Limits for the Ratio of Means'. Symposium on Interval Estimation, *J. Royal Stat. Soc., Ser. B*, 16–17 (1954–55), 186–94.

Creasy, M.A., 'Confidence Limits for the Gradient in the Linear Functional Relationship', *J. Royal Stat. Soc., Ser. B* 18 (1956), 65–9.

Dawid, A.P., 'On the Concepts of Sufficiency and Ancillarity in the Presence of Nuisance Parameters', *J. Royal Stat. Soc., Ser. B* 36 (1975), 248–58.

Dawid, A.P. and Stone, M., 'Un-Bayesian Implications of Improper Bayes Inferences in Routine Statistical Problems', *Biometrika* 59 (1972), 369–75.

Dawid, A.P., Stone, M. and Zidek, J.V., 'Marginalization Paradoxes in Bayesian and Structural Inference', *J. Royal Stat. Soc., Ser. B* 35 (1973), 189–233.

DeFinetti, Bruno, *Probability, Induction and Statistics.* London: John Wiley and Sons, 1972.

DeFinetti, Bruno, 'Bayesianism: Its Unifying Role for Both the Foundations and the Applications of Statistics', *Proc. 39th Int. Stat. Inst.*, #16.2 (1973), 349–76.

Dempster, A.P., 'On Direct Probabilities', *J. Royal Stat. Soc., Ser. B* 25 (1963), 100–10.

Dempster, A.P., 'Further Examples of Inconsistencies in the Fiducial Argument', *Annals of Math. Stat.* 34 (1963), 884–91.

Dempster, A.P., 'On the Difficulties Inherent in Fisher's Fiducial Argument', *J. Am. Stat. Assoc.* 59 (1964), 56–66.

Dempster, A.P., 'New Methods for Reasoning Towards Posterior Distributions Based on Sample Data', *Annals of Math. Stat.* 37 (1966), 355–74.

Dempster, A.P., 'Review of: *Logic of Statistical Inference*', *J. Am. Stat. Assoc.* 61 (1966), 1233–5.

Dubins, L., 'Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations', *Annals of Probability* 3, #1 (1975), 89–99.

Edwards, A.W.F., *Likelihood.* Cambridge: Cambridge Univ. Press, 1972.

Fieller, E.C., 'Some Problems in Interval Estimation', Symposium on Interval Estimation, *J. Royal Stat. Soc., Ser. B* 16–17 (1954–55), 175–85.

Finney, D.J., *An Introduction to the Theory of Experimental Design.* Chicago: Univ. of Chicago, 1963.

Fisher, R.A., *Contributions to Mathematical Statistics.* New York: John Wiley & Sons, Inc., 1950.

Fisher, R.A., *The Design of Experiments*, 9th ed. New York: Hafner, 1971.

Fisher, R.A., *Statistical Methods for Research Workers*, 14th ed. New York: Hafner, 1973.

Fisher, R.A., *Statistical Methods and Scientific Inference*, 3rd ed. New York: Hafner, 1973.

Fisher, R.A., *Collected Papers of R.A. Fisher.* J.H. Bennett ed., 5 volumes. Adelaide: Univ. of Adelaide, 1971–1975.

Fraser, D.A.S., 'On Fiducial Inference', *Annals Math. Stat.* 32, #3 (1961), 661–76.

Fraser, D.A.S., *The Structure of Inference.* New York: John Wiley and Sons, 1968.

Geisser, S., 'Book Review of *Introduction to Probability and Statistics from a Bayesian Viewpoint*', *Technometrics* 8, #2 (1966).

Ghosh, J.K., 'On the Relation Among Shortest Confidence Intervals of Different Types', *Calcutta Stat. Association Bull.* 10 (1961), 147–152.

Ghosh, J.K. and Subramanyam, K., 'Second Order Efficiency of Maximum Likelihood Estimators', *Sankhyā* 36 (1974), 325–58.

Gillies, D.A., *An Objective Theory of Probability.* London: Methuen, 1973.

Good, I.J., *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge: The M.I.T. Press, 1965.

Good, I.J., 'Random Thoughts about Randomness', in *PSA 1972.* K. Schaffner and R. Cohen eds., Dordrecht: D. Reidel, 1974.

Hacking, Ian, *Logic of Statistical Inference.* Cambridge: Cambridge Univ. Press, 1965.

Hilpinen, R., 'On the Information Provided by Observations', in Hintikka, J. and Suppes, P. eds, *Information and Inference*. Boston and Dordrecht: D. Reidel, 1970.
Jeffreys, Harold, *Theory of Probability*, 3rd ed. Oxford: Clarendon Press, 1967.
Kemeny, J., 'Fair Bets and Inductive Probabilities', *Journal of Symbolic Logic* 20, #3 (1955), 263–273.
Keynes, J., *Treatise of Probability*, 2nd ed. New York: Macmillan and Co., 1921.
Kolmogorov, A.N., *Foundations of the Theory of Probability*, 2nd ed. New York: Chelsea Publishing Company, 1956.
Kyburg, H.E. Jr., *Probability and the Logic of Rational Belief*. Middletown: Wesleyan Univ. Press, 1961.
Kyburg, H.E. Jr. and Smokler, H.E. eds., *Studies in Subjective Probability*. New York: John Wiley & Sons, Inc., 1964.
Kyburg, H.E. Jr., *The Logical Foundations of Statistical Inference*. Boston and Dordrecht: D. Reidel, 1974.
Laplace, S., *A Philosophical Essay on Probabilities*, trans. and rep. New York: Dover, 1952.
Lehman, R.S., 'On Confirmation and Rational Betting', *Journal of Symbolic Logic* 20, #3 (1955), 251–262.
Levi, Isaac, 'On Indeterminate Probabilities', *J. of Phil.* LXXI (1974), 391–417.
Levi, Isaac, 'Direct Inference', *Journal of Philosophy*, LXXIV, #1 (1977), 5–29.
Lindley, D.V., 'Review of: *Statistical Methods and Scientific Inference*', *Heredity* 11 (1957), 280–3.
Lindley, D.V., 'Fiducial Distributions and Bayes' Theorem', *J. Royal. Stat. Soc., Ser. B* 20–21 (1958), 102–7.
Lindley, D.V., *Introduction to Probability and Statistics from a Bayesian Viewpoint*. 2 volumes, Cambridge: Cambridge Univ. Press, 1969.
Lindley, D.V. and El-Sayyad, G.M., 'The Bayesian Estimation of a Linear Functional Relationship', *J. Royal Stat. Soc., Ser B* 30 (1969), 190–202.
Mauldon, J.G., 'Pivotal Quantities for Wishart's and Related Distributions, and a Paradox in Fiducial Theory', *J. Royal Stat. Soc., Ser. B* 17 (1955), 79–85.
McKinlay, S.M., 'The Design and Analysis of the Observational Study – A Review', *J. Am. Stat. Assoc.* 70, #351 (1975), 503–520.
Nagel, Ernest, *Principles of the Theory of Probability*. volume 1, #6 International Encyclopedia of Unified Science. Chicago: Univ. of Chicago Press, 1969.
Neyman, J., ' "Inductive Behavior" as a Basic Concept of Philosophy of Science', *Bull. Int. Stat. Inst.* 25 (1957), 7–22.
Neyman, J., *A Selection of Early Statistical Papers of J. Neyman*. Berkeley: Univ. of California Press, 1967.
Neyman, J. and Pearson, E.S., *Joint Statistical Papers*. Berkeley: Univ. of California Press, 1967.
Pearson, E.S., *The Selected Papers of E.S. Pearson*. Berkeley: Univ. of California Press, 1966.
Rao, C.R., Chairman: Conference on Fiducial Probability, *Bull. Int. Stat. Inst.* 40[2] (1963), 833–939.
Redhead, M.L.G., 'On Neyman's Paradox and the Theory of Statistical Tests', *Brit. J. Phil. Sci.* 25 (1974), 265–271.
Renyi, A., 'On a New Axiomatic Theory of Probability', *Acta Math. Acad. Sci.*

Hungaricae 6 (1955), 285–335.
Renyi, A., *Foundations of Probability*. San Francisco: Holden-Day, 1970.
Robbins, H., 'A New Approach to a Classical Statistical Decision Problem' in Kyburg, H. and Nagel, E. eds., *Induction: Some Current Issues*. Wesleyan Univ. Press, 1963.
Savage, L.J., *The Foundations of Statistics*. New York: Dover, 1972.
Savage, L.J., 'On Rereading R.A. Fisher', *The Annals of Statistics* 4, #3 (1976), 441–500.
Shimony, A., 'Coherence and the Axioms of Confirmation', *Journal of Symbolic Logic* 20, #1 (1955), 1–28.
Spielman, S., 'On the Infirmities of Gillies' Rule', *Brit. Journal Phil. Sci.* 25 (1974), 261–65.
Sprott, D.A., 'Transformations and Sufficiency', *J. Royal Stat. Soc., Ser B* 27 (1965), 479–85.
Sprott, D.A. and Kalbfleisch, J.G., 'Fiducial Probability', *Statistische Hefte* 2 (1967), 99–109.
Sprott, D.A. and Kalbfleisch, J.D., 'Application of Likelihood Methods to Models Involving Large Numbers of Parameters', *J. Royal Stat. Soc., Ser B* 32 (1970), 175–208.
Sprott, D.A. and Kalbfleisch, J.D., 'Marginal and Conditional Likelihoods', *Sankyhā, Ser. A* 35 (1973), 311–28.
Stein, C., 'An Example of Wide Discrepancy Between Fiducial and Confidence Intervals', *Annals Math. Stat.* 30 (1959), 877–80.
Stein, Charles, 'A Remark on the Likelihood Principle', *J. Royal Stat. Soc., Ser. A* 125 (1962), 565–8.
'Student', 'Comparison Between Balanced and Random Arrangements of Field Plots', *Biometrika* 39 (1937), 363–79.
Tukey, J.W., 'Some Examples with Fiducial Relevance', *Annals of Math. Stat.* 28 (1957), 687–95.
Verhagen, A.M.W., *The Notion of Induced Probability in Statistical Inference*. Technical Paper #21, Melbourne: CSIRO, 1966.
Venn, J., *The Logic of Chance*, 4th ed. New York: Chelsea Publishing Company, 1962.
Wald, A., *On the Principles of Statistical Inference*. Notre Dame Mathematical Lectures #1, Notre Dame: Univ. of Notre Dame, 1942.
Wald, A., *Wald: Selected Papers in Statistics and Probability*. Stanford: Stanford Univ. Press, 1957.
Wald, A., *Statistical Decision Functions*. New York: Chelsea Publishing Company, 1971.
Welch, B.L., 'On Confidence Regions and Sufficiency, with Particular Reference to Parameters of Location', *Annals Math. Stat.* 10 (1939), 59–69.
Wilks, S.S., 'Shortest Average Confidence Intervals from Large Samples', *Annals of Math. Stat.* 9 (1938), 166–75.
Wilks, S.S., 'The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses', *Annals of Math. Stat.* 9 (1938), 60–2.
Williams, E.J., 'A Comparison of the Direct and Fiducial Arguments in the Estimation of a Parameter', *J. Royal Stat. Soc., Ser B* 25 (1963), 95–9.
Williams, J.S., 'The Role of Probability in Fiducial Inference', *Sankhyā* 28 (1966), 271–96.
Yates, F., 'Comments on D.V. Lindley's Review of Sir Ronald Fisher's *Statistical Methods and Scientific Inference*', *Heredity* 12 (1958), 133–5.
Yule, G.U. and Kendall, M.G., *An Introduction to the Theory of Statistics*, 12th ed. London: 1940.

INDEX

(compiled, using *BINDEX*, with thanks to Nuel Belnap)

Anderson, T.W. 232, 238
Anscombe, F.J. 100

Barnard, G. 15, 26, 102, 164
Bartlett, M.S. 164
Bayes' theorem 7, 11, 23
Bayes, T. 11, 26
Behrens-Fisher problem 63–64, 148–49
Bennett, G.W. 165
Birnbaum, A. 118, 119, 134
Boole, G. 11, 26
Braithwaite, R.B. 100
Brillinger, D.R. 156, 157, 158, 165, 166, 177
Buehler, R.J. 163, 166

calculable transformation 89, 155–57
Carnap, R. 6, 17, 25, 26
Clopper, C.J. 134
Cohen, R. 224
conditional density 22–23
confirmation 6–14
  coherence 7
  conditionalization 6–8
  frequency principle 9, 172, 231–32
  insufficient reason 10–14, 23
  objectivity 10
  relevance 16–21, 90–95
  relevance and experimental design 202–25
  total evidence 8, 83
Cornish, E.A. 165, 224, 238
countable additivity, *See* de Finetti, B.; Dubins, L.; fiducial inference, conglomerability; improper distribution; Renyi, A.
Cox, D.R. 28, 44, 66
Cramer, H. 68, 100, 101, 206, 223
Creasy, M. 61, 69, 165
Creasy-Fieller paradox 60–63, 154–55, 236–38
credence 6
credibility 6

Dawid, A.P. 140, 156, 159, 161, 162, 164, 166
de Finetti, B. 2, 24, 163, 226, 237
Dempster, A.P. 65, 156, 159, 160, 162, 166, 177
density function 22
direct inference 2, 107–9, 169–71
  Fisherian theory 71–76, 186–87
  Hacking's program 172
  Kyburg's program 184–89
  Levi's criticism of Kyburg's program 200
  pivotal 266–27
  reference class 31–33, 65, 73–75, 75, 108–9, 118, 184–89, 191
distribution function 21
Dubins, L. 167

Edwards, A.W.F. 19, 26, 118, 134
epistemological probability. *See* direct inference, Kyburg's program; fiducial inference, Kyburg's program
estimation
  consistent statistic 205, 222–23
  efficiency 91–93, 206, 223–24
  and experimental design 205–14
  and Information 204–8, 223–24
  and relevance 90–93, 205–8
  sufficient statistic 206–7

Feddersen, A.P. 163, 166
Feller, W. 24, 27
fiducial inference 4, 103–67
  compatibility with conditionalization 120–24, 126, 129–34, 228–30. *See also* Lindley, D.V.

compatibility with significance tests 126–29, 147, 152
conglomerability 163–64, 167, 226–27
exhaustive statistic 112–16. *See also* statistic, exhaustive
fiducial priors 116–17, 139–40
group invariance 166
Hacking's example 108
Hacking's program 168–78
Kyburg's program 189–95
Lindley's theorem 122–25
objections to Hacking's program 176–78
objections to Kyburg's program 192–95
one parameter 103–35
paradoxes 131, 145–46, 151–64
pivotal 107–12, 125–26, 226–27. *See also* pivotal
several parameters 136–51
step-by-step method 138–40
Fieller, E.C. 37, 60, 69, 155, 159, 236, 238
Fisher, R.A. 3, 5, 14, 19, 24, 26, 27, 46, 66, 67, 69, 71, 72, 73, 74, 84, 85, 86, 90, 91, 96, 100, 101, 105, 106, 109, 111, 117, 121, 122, 123, 130, 133, 134, 135, 137, 140, 144, 145, 146, 151, 153, 155, 159, 164, 165, 171, 178, 196, 199, 200, 204, 205, 206, 218, 222, 223, 224, 225, 227, 238
Fraser, D.A.S. 26, 68, 165, 229

Geisser, S. 165
Ghosh, J.K. 68, 223
Gillies, D.A. 87, 92, 101
Good, I.J. 130, 135, 155, 156, 212, 224
Gossett, W.S. 144, 218, 225
group invariance. *See* fiducial inference, group invariance; Fraser, D.A.S.

Hacking, I. 4, 5, 9, 15, 25, 26, 28, 30, 36, 37, 47, 48, 49, 66, 67, 107, 129, 134, 140, 169, 171, 172, 173, 174, 176, 177, 179, 180, 181, 182, 192, 220, 229, 230, 231, 238
Hilpinen, R. 222

improper distribution 23–24, 160–64. *See also* de Finetti, B.; Dubins, L.; fiducial inference, conglomerability; Renyi, A.
inverse inference
bayesian 2, 11, 18–19
frequentist 2–3, 75, 171, 178–81, 189–92

Jeffrey, R. 25
Jeffreys, H. 12, 26, 76, 84, 86, 100, 101, 140, 164, 165
joint distribution 22

Kalbfleisch J.D. 164
Kemeny, J. 25
Kendall, M.G. 84, 100
Keynes, J. 26
Kolmogorov, A.N. 25, 135
Kyburg, H.E. 5, 24, 25, 44, 64, 66, 67, 75, 100, 181, 183, 186, 187, 188, 189, 191, 192, 197, 198, 199, 200, 220, 225

Laplace, P. 10, 25, 26
Lehman, R. 25
Levi, I. 5, 7, 25, 100, 166, 183, 184, 192, 199, 200
likelihood
Hacking's experimental densities 175–77
Hacking's irrelevance principle 173
Hacking's law of 170–71
likelihood function 19
estimation 206–7
and significance tests 80, 96, 128–29
Lindley, D.V. 3, 24, 25, 26, 27, 65, 120, 121, 122, 123, 124, 131, 134, 135, 145, 146, 155, 156, 158, 161, 162, 165, 166, 175, 178, 228
location parameter 24, 80, 123, 128

marginal distribution 22
master significance test [MST] 59–63, 154–55

Maukon, J.G. 156, 157, 165
McKinlay, S.M. 222
Mendel, G. 85
Morgenbesser, S. 134
Morrison, N. 135

Nagel, E. 65
Neyman, J. 3, 15, 24, 26, 27, 28, 29, 30, 32, 33, 34, 35, 38, 46, 51, 53, 64, 65, 66, 67, 68, 69, 87, 92, 93, 94, 100, 101, 153
N-P theory 4, 14–16, 28–69
  confidence equivalence 59
  confidence intervals 51–58, 152–54
  criticism of confidence intervals 56–58
  direct inference 30–36
  Hacking's criticism 15, 36–37, 47–49
  hypothesis tests 37–47
  inductive behavior 39
  likelihood ratio test 42, 44, 45–49, 50, 55
  mixed tests 42–45, 153
  N-P lemma 42
  stopping rule problem 49
  type 1 error 40
  type 2 error 40

Pearson, E.S. 27, 29, 65, 66, 67, 134
pivotal 226–27. *See also* fiducial inference, pivotal
  canonical 125, 228–30
  smoothly invertible 100–111, 174, 177, 227, 230–31
Pratt, J.W. 68

random sampling 178–81, 195–99, 214–22
Rao, C.R. 223
Redhead, M.L.G. 87, 89, 101
Reichenbach, H. 9, 25, 28, 31, 32, 65, 134, 223
Renyi, A. 23, 25, 27
Robbins, H. 30, 65

Savage, L.J. 65, 67, 107, 134, 135, 156, 158, 222, 223
scale parameter 24, 123
Schaffner, K. 224
Shimony, A. 25
Significance tests 76–102
  chi-squared 84–86
  contingency tables 95–99
  criticisms 80
  discrepancy ranking 78–80
  heuristic analogy 81–82
  lottery paradox 82
  Neyman's criticism 87, 92–95
Smokler, H. 24, 67
Spielman, S. 87, 89, 101
Sprott, D.A. 164
statistic 19
  ancillary 20–21, 97–99 112–16, 119–20, 152
  exhaustive 112–16, 152
  sufficient 20, 83, 112–16, 138, 152, 192–95. *See also* estimation
Stein, C. 167
Stone, M. 140, 156, 159, 161, 162, 164, 166
Subramanyam, K. 223
Suppes, P. 134

Tukey, J.W. 111, 134, 156, 165, 166, 174, 182, 227, 230

Venn, J. 11, 26
Verhagen, A.M. 224, 238
von Kries 26
von Mises, R. 9, 25

Wald, A. 30, 46, 65, 67, 68
Welch, B.L. 68, 152, 153, 165
White, M. 134
Wilks, S.S. 68
Williams, J.S. 238

Yates, F. 146, 165
Yule, G.U. 84, 100

Zidek, J.V. 140, 156, 159, 161, 164, 166

THEORY AND DECISION LIBRARY

An International Series in the Philosophy and Methodology of the Social and Behavioral Sciences

Editors:

Gerald Eberlein, *University of Technology, Munich*
Werner Leinfellner, *University of Nebraska*

1. Günther Menges (ed.), *Information, Inference, and Decision*. 1974, viii + 195 pp.
2. Anatol Rapoport (ed.), *Game Theory as a Theory of Conflict Resolution*. 1974, v + 283 pp.
3. Mario Bunge (ed.), *The Methodological Unity of Science*. 1973, viii + 264 pp.
4. Colin Cherry (ed.), *Pragmatic Aspects of Human Communication*. 1974, ix + 178 pp.
5. Friedrich Rapp (ed.), *Contributions to a Philosophy of Technology. Studies in the Structure of Thinking in the Technological Sciences*. 1974, xv + 228 pp.
6. Werner Leinfellner and Eckehart Köhler (eds.), *Developments in the Methodology of Social Science*. 1974, x + 430 pp.
7. Jacob Marschak, *Economic Information, Decision and Prediction. Selected Essays*. 1974, three volumes, xviii + 389 pp.; xii + 362 pp.; x + 399 pp.
8. Carl-Axel S. Staël von Holstein (ed.), *The Concept of Probability in Psychological Experiments*. 1974, xi + 153 pp.
9. Heinz J. Skala, *Non-Archimedean Utility Theory*. 1975, xii + 138 pp.
10. Karin D. Knorr, Hermann Strasser, and Hans Georg Zilian (eds.), *Determinants and Controls of Scientific Developments*. 1975, ix + 460 pp.
11. Dirk Wendt, and Charles Vlek (eds.), *Utility, Probability, and Human Decision Making. Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973*. 1975, viii + 418 pp.
12. John C. Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation*. 1976, xvi + 262 pp.
13. Gerhard Schwödiauer (ed.), *Equilibrium and Disequilibrium in Economic Theory. Proceedings of a Conference Organized by the Institute for Advanced Studies, Vienna, Austria, July 3–5, 1974*. 1978, l + 736 pp.
14. V. V. Kolbin, *Stochastic Programming*. 1977, xii + 195 pp.
15. R. Mattessich, *Instrumental Reasoning and Systems Methodology*. 1978, xxii + 396 pp.
16. H. Jungermann and G. de Zeeuw (eds.), *Decision Making and Change in Human Affairs*. 1977, xv + 526 pp.
18. A. Rapoport, W. E. Stein, and G. J. Burkheimer, *Response Models for Detection of Change*. 1978 (forthcoming)
19. H. J. Johnson, J. J. Leach, and R. G. Mühlmann (eds.), *Revolutions, Systems, and Theories; Essays in Political Philosophy*. 1978
20. Stephen Gale and Gunnar Olsson (eds.), *Philosophy in Geography*. 1979, xxii + 470 pp.
21. Maurice Allais and Ole Hagen (eds.), *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of Decisions Under Uncertainty With Allais' Rejoinder*. 1979, vii + 681 pp. + indexes.
22. Teddy Seidenfeld, *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher*. 1979, xiv + 246 pp.

709653